

VESTAL, RICHARD D., M.S., Sequence Variation in *LEAFY*, a Candidate Gene for Life History Variation in *Arabidopsis lyrata*. (2010)  
Directed by Dr. David L. Remington. 75 pp.

*LEAFY* (*LFY*) is a functional candidate for variation in resource allocation in *Arabidopsis lyrata*. We isolated and characterized *LFY* alleles from *A. lyrata* individuals from Spiterstulen Norway, Plech Germany, Ithaca N.Y., USA, and Mayodan N.C, USA. We found numerous coding polymorphisms, insertions, and deletions that were only in certain populations or locations with the majority of the variation in the European populations. Our data supports the idea that Central Europe has served as refugia for *A. lyrata*. We identified two possible miRNAs, one in all North American individuals and one in some European individuals. We also found a Little Athila retro-element in a European population. Our population genetics analyses found evidence that the majority of the polymorphisms in *A. lyrata* are deleterious and have not been eliminated by natural selection. Our research suggests that the power of genetic drift in small, out-crossing populations has a similar effect to inbreeding in larger populations. Deleterious mutations that would normally be eliminated by natural selection are kept at high frequencies and may reach fixation at a faster rate.

SEQUENCE VARIATION IN *LEAFY*, A CANDIDATE GENE FOR  
LIFE HISTORY VARIATION IN *ARABIDOPSIS LYRATA*

by

Richard D. Vestal

A Thesis Submitted to  
the Faculty of The Graduate School at  
The University of North Carolina at Greensboro  
in Partial Fulfillment  
of the Requirements for the Degree  
Master of Science

Greensboro  
2010

Approved by

David L. Remington

Committee Chair

## APPROVAL PAGE

This thesis has been approved by the following committee of the Faculty of The  
Graduate School at The University of North Carolina at Greensboro

Committee Chair \_\_\_\_\_  
David L. Remington

Committee Members \_\_\_\_\_  
Karen Katula

\_\_\_\_\_  
Malcolm Schug

\_\_\_\_\_  
Date of Acceptance by Committee

\_\_\_\_\_  
Date of Final Oral Examination

## **ACKNOWLEDGMENTS**

I thank Dr. David L. Remington for being my advisor and giving me the opportunity to be on this project. I also thank Dr. David L. Remington for his guidance and support throughout this project and for giving me the opportunity to work as a genetics researcher.

I would like to thank the Henrich Lab, especially Josh, for all of their assistance and guidance in the analysis of my preliminary data and usage of the MegaBACE sequencer. I would also like to thank my other committee members: Dr. Karen Katula and Dr. Malcolm Schug for providing comments, insights, and suggestions to improve my thesis. I would like to thank Dr. Malcolm Schug for his assistance analyzing the population genetics statistics.

Funding was provided by the UNCG Graduate Research Fund and the Biology Department at UNCG. Additional funding was provided by Dr. David Remington's Lab

Importantly, I thank my wife Ann Vestal for her support and encouragement through this process.

## TABLE OF CONTENTS

	Page
CHAPTER	
I. INTRODUCTION .....	1
II. METHODS .....	24
III. RESULTS.....	30
IV. DISCUSSION .....	46
LITERATURE CITED .....	63
APPENDIX A.1: Region lengths in Sequence Standards and Alleles Amplified.....	71
APPENDIX A.2: Region lengths in Sequence Standards and Alleles Amplified- cont'd .....	72
APPENDIX A.3: Amino acids 1-140 for all sequences .....	73
APPENDIX A.4: Amino acids 141-280 for all sequences .....	74
APPENDIX A.5: Amino acids 281-422 for all sequences .....	75

## CHAPTER I

### INTRODUCTION

*LEAFY (LFY)* is a functional candidate for variation in resource allocation in *Arabidopsis lyrata*. Researchers have previously shown the presence of an 88 base pair (bp) repeat insertion in the intron1 region of *LFY* and variation in intronic regions can lead to alterations in gene expression. Using the *Arabidopsis thaliana* sequence for *LFY* we initially designed PCR primer sequences specific to introns 1 and 2 of *LFY* and used these to identify polymorphisms of *LFY* alleles in individuals in populations from Plech Germany and Mayodan North Carolina, USA. This was the first step towards characterizing the variation in sequence and gene expression in *A. lyrata* populations with different resource allocation phenotypes.

In the life history of plants, two of the major phase changes are when to grow and when to reproduce or flower. Life history strategies with respect to reproduction can be classified in two ways, iteroparous or semelparous. Semelparous species reproduce once and then die like *A. thaliana*. Iteroparous species, like *A. thaliana*'s closest relative *A. lyrata*, reproduce several times over multiple years. It is important to note that iteroparous organisms can vary in the degree and pattern to which they allocate resources

towards reproduction vs. vegetative growth. In order for the change from a vegetative state to a reproductive state to take place, meristematic cells or undifferentiated plant cells, must receive the proper stimulus to initiate flowering and homozygous *lfy* mutants are sterile (Bonser and Aarssen 2006). These two life history strategies allocate the available environmental resources differently based on their varied requirements. The populations chosen for this study cover the entire *A. lyrata* growing range from the northern to southern extremes and vary greatly in phenotype.

*LFY* is a logical candidate for modulating resource allocation patterns given its role as the primary regulator of the transition from a vegetative state to a reproductive state. Variation in *LFY*, including coding sequence, introns, and flanking regions is responsible for the timing, levels, and patterns of *LFY* gene expression. Variation in the coding sequence can change the binding affinity of the LFY protein and may impair its ability to activate downstream targets. Variation in the non-coding sequences can alter the ability of upstream activators to bind effectively to *LEAFY* and activate transcription (Blázquez et. al. 1997).

My specific aims are to identify and characterize *LFY* alleles from multiple individuals across the *A. lyrata* growing range in order to find evidence of selective pressure being applied to *LFY*. Our methods will include primers designed to isolate the entire *LFY* gene region including flanking sequences, the creation of consensus sequences, functional genetics analysis including characterization of conserved domains and structural elements and the identification of known transcription factor binding site

motifs, and population genetics analysis including the MacDonald Kreitman tests and the Tajima's D test.

The goal of this research project will be to evaluate the evidence for variation in *LFY*'s coding, non-coding, and flanking regions that may be consistent with involvement in local adaptation due to pressure from natural selection. The two primary hypotheses and sub-hypotheses are as follows:

(1) I hypothesize that DNA sequence differences in *LFY* are present that are consistent with a functional role in the differences in the respective semelparous versus iteroparous life histories between *A. lyrata* and *A. thaliana*.

(2) I hypothesize that *LFY* polymorphisms are present that are consistent with a functional role in the differences in the resource allocation strategies between *A. lyrata* populations.

## **Background**

Semelparous species, such as *A. thaliana*, will make a major resource acquisition commitment to initial growth and biomass increase before making the one all-encompassing switch to flowering as it leads directly to the death of the parent plant. Iteroparous species are not restricted in such a manner and can better delegate the actual allocation of resources depending on environmental conditions and resource availability. Reproductive effort may be reduced in resource limited years, due to events like drought, prolonged periods of poor light quality after fires and volcanic eruptions, or harsh



weather, if it may lead to increased reproductive ability in the year or years to come.

There can also be variation in “hard-wired” allocation patterns in different populations, possibly due to selection in different environments (Doust 1989).

*A. thaliana* belongs to the family Brassicaceae and is native to Asia, Northwestern Africa, and Europe. This self pollinating annual plant grows about ten inches high with a rosette of leaves at the base. The rosette is a vegetative shoot with no internode elongation. The plant has a primary stem that flowers and has few leaves as shown in Figure 1.



Figure 1 *A. thaliana* Page and Grossniklaus 2002  
doi:10.1038/nrg730



Figure 2 *A. lyrata*  
<http://www.jgi.doe.gov/sequencing/why/3066.html>

*A. thaliana* will eventually flower in days with short light periods, but flowers much more rapidly when exposed to long periods of daylight and needs a transient exposure to cold weather, vernalization, to initiate flowering. If these conditions are not met then the plant can eventually flower through an age-related pathway (Ehrenreich and Purugganan 2006). This model plant for genetics has had its entire genome of five chromosomes sequenced. The selfing nature, rapid life cycle, and ability to produce a large quantity of siliques each containing about 20-30- seeds makes this a very useful tool for genetics research (Al-Shehbaz and O’Kane 2002).

*A. lyrata* belongs to a lineage that diverged from *A. thaliana* roughly five million years ago. It has a complement of eight chromosomes instead of five like *A. thaliana*. *A. lyrata* is a non self-pollinating and generally self-incompatible perennial that prefers altitudes of less than 1500 meters, environments that are rocky, sandy, or freshwater shoreline, both in North America (ssp. *lyrata*) and Eurasia (ssp. *petraea*) (Koch and Matschinger 2007). *A. thaliana* probably has axillary meristems, but they remain quiescent. In *A. lyrata* they can be active and develop into vegetative or reproductive shoots as seen in Figure 2. The similarities of the two species make it possible to transfer many of the mechanisms involved in the flowering pathways from *A. thaliana* to *A. lyrata*. The primary mechanism addressed here is the role of DNA sequence variation in *LFY* and how it is represented in phenotypic variation among the sequenced *A. lyrata* sample populations.

Plants have meristematic tissues in the roots and shoots of the plant where growth will take place. The continued replenishment of these cells allows some cells to differentiate into new structures including floral organs and allows others to enable the plant to grow in

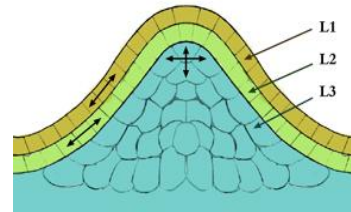


Figure 3 Tunica-Corpus model Epidermal (L1) and subepidermal (L2) layers form the tunica. The inner L3 layer is called the corpus. Wikimedia commons

length and girth or respond to trauma. Apical meristems at the root or shoot apex consist solely of undifferentiated cells. These regions will allow for indeterminate growth and are responsible for producing the three types of primary meristems.

These three types are from outside the plant to inside: (1) the Protoderm that becomes the epidermis: (2) the Procambium is next internally and will become the primary xylem and primary phloem. In angiosperms other than monocots it will also make one secondary lateral meristem, the vascular cambium responsible for secondary xylem and secondary phloem: and (3) the Ground meristem that becomes the pith. This region is also responsible for forming the other secondary or lateral meristem, the cork cambium. The lateral meristems are involved in lateral growth. Apical meristems are divided by layers and zones in order to control growth, see Figures 3 and 4 (Blázquez et al 2006).

Once the shoot apical meristem (SAM) has formed, the required level of stem cells for organogenesis is maintained by the *CLAVATA* (*CLV1*, *CLV2*, and *CLV3*) gene family.

Excess stem cells will accumulate in the center of the SAM if *CLV* loci function is lost. *CLV1*

produces an extracellular Leu-rich repeat (LRR) receptor Serine/Threonine kinase while the *CLV2* gene forms an LRR receptor-like protein, and *CLV3* manufactures a small secreted protein of the plant specific CLE family. *CLV3* is found in regions L1 and L2 while *CLV1* is located in regions L2 and L3 of the SAM as seen in Figure 5 (Sharma and Fletcher 2002).

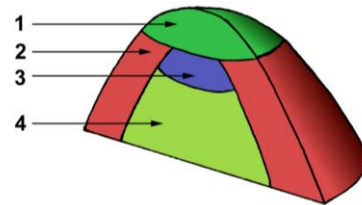


Figure 4 Organization of an apical meristem 1-Central zone 2-Peripheral zone 3-Medullary (central) meristem 4-Medullary tissue. Wikimedia commons

The shoot apical meristem (SAM) is responsible for all above ground growth and is responsible for producing

axillary stems or branches, leaves, and the organs of the flowers. The root apical meristem or (RAM) is

responsible for below ground functions. If a meristem is

vegetative, the SAM will manufacture the leaves and shoots. Inflorescence meristems,

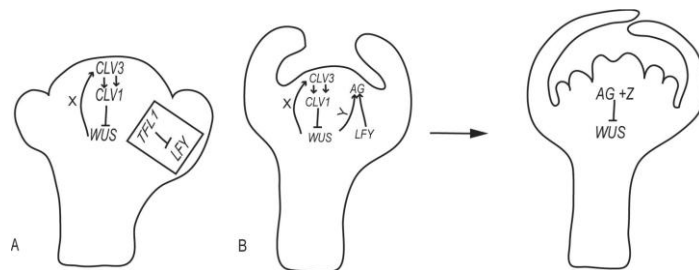


Figure 5 Model interactions between regulatory genes in SAM & floral meristem. Sharma and Fletcher 2002

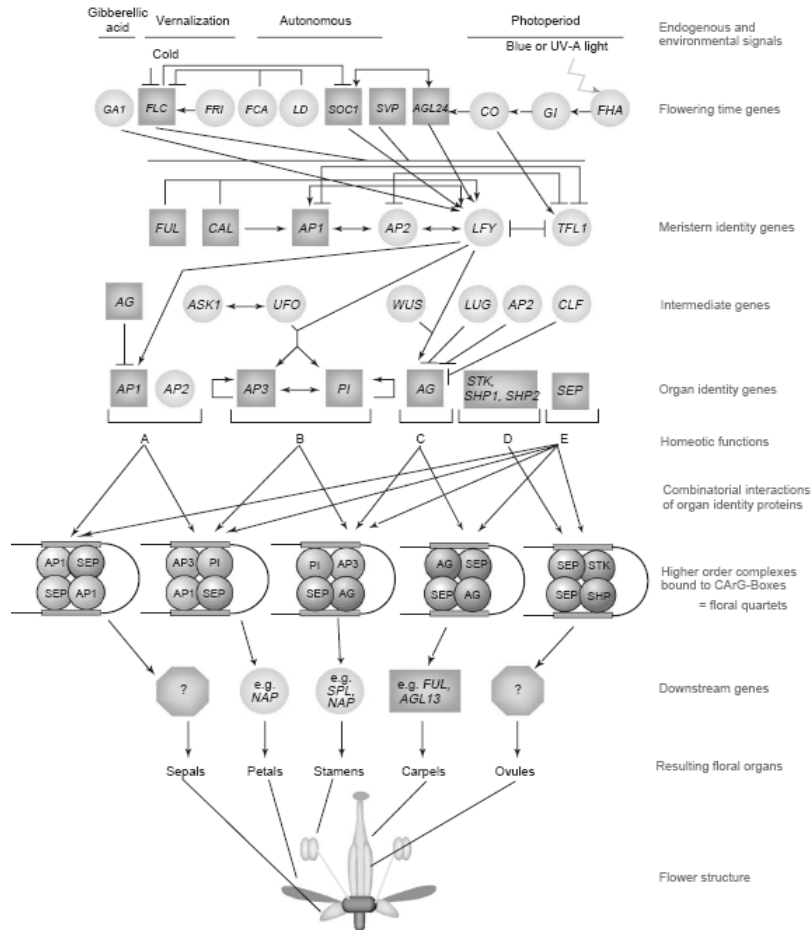


Figure 6 the Four Floral Induction Pathways of *LEAFY*-  
Kaufmann, K., Melzer, R., and Theißen, G. (2005)

the transition state, will

produce floral meristems that will be responsible for producing the sepals, petals, stamens, and carpels (Sharma and Fletcher 2002). These organs develop in concentric rings or whorls unless there is a mutation in one or more of the homeotic genes controlling their development. After the whorls develop activity ceases in the floral

meristem unlike the SAM which stays indeterminate and continues to grow (Weigel 1995, Shepard 2007).

In *Arabidopsis* several genes have been identified that are necessary for proper SAM formation and maintenance. Some of the main genes involved *CUP-SHAPED COTYLEDON 1* and 2 (*CUC1* and 2), *SHOOTMERISTEMLESS (STM)* and *WUSCHEL*. Once the commitment to become a floral meristem is decided then *LFY*, along with co-regulators such as *WUS* for *AG* or the UFO F-Box protein for *AP3* and *PI*, will regulate the transcription of the floral homeotic genes *API*, *AP3*, and *AG* that are responsible for the proper positioning of the floral organs as seen in Figure 6.

*LFY* has a unique sequence not found in other proteins making it the sole member of its gene family. *LFY's* importance is exemplified by the conservation of its DNA sequence in taxa varying from mosses to angiosperms (Maizel et al, 2005). The *lfy* mutant phenotype consisting of total or incomplete conversion of floral meristems to shoots has been observed in other plants like maize, petunias, snapdragons, and tomatoes. Typically these plants have early-arising flowers that are completely transformed into inflorescence shoots and the late-arising are partially transformed as seen in Figure 7 (ABRC). They have no petals or stamens and are sterile. The severest phenotypic reaction is seen under short days and lower temperatures. In gymnosperms, a *LFY* paralog *NEEDLY (NLY)* has been found; however no mutant phenotype is available. *LFY* and *NLY* have been shown to have similar expression profiles. An interesting point to

note is that *NLY* disappeared with the arrival of flowers, putting the focus of successful proliferation of the angiosperms squarely on *LFY*.

The *LFY* transcription factor is known to function independently as a homodimer and also in complexes with *WUSCHEL* (*WUS*) and *UNUSUAL FLORAL ORGANS* (*UFO*). It is able to induce its own transcription by recognizing and binding to its own target sequence consisting of the (CCANTGT/G) motif in gene promoter regions. *API* is known to have one such site and *AG* has four sites numbered *AG-I* to *AG-IV*. (Parcy, Bomblies, and Weigel 2002, Busch, Bomblies, and Weigel 1999). Due to *LFY*'s continuous expression

throughout the plants lifetime, this same motif will be searched for in all sequences as a means to verify whether or not *LFY* has the means to directly self-regulate its own transcription. *LFY* RNA levels have been detected in 9 day old seedlings and initially begin to increase in floral primordial and the flanks of the Inflorescence Meristem. *LFY* then moves from the inflorescence meristem to the floral meristem as it develops. As the Floral meristem develops, *LFY* RNA is found in young flowers during stages 1/2 but is not found in the center during stages 3/4 (Cary et al 2002).



Figure 7 *lfy-11* strong mutant phenotype

*LFY* has two primary conserved regions, the N- and C-Domains. The C-terminal region of *LFY* is responsible for DNA binding and the N-terminal is involved in transcriptional regulation (Maizel et al 2005). Sequence analysis has shown that some regions of *LFY* are strictly conserved amongst all species of plants however some regions are specific to angiosperms alone. The importance in sequence variation can be seen by the fact that R390, a pivotal residue involved in monomer formation and cooperative binding interactions, is found in all angiosperm *LFY* proteins. This arginine has been substituted with a lysine in gymnosperms and ferns. This single amino-acid change from lysine to arginine is responsible for reducing *LFY*'s DNA binding affinity by lowering the bond strength between the *LFY* protein monomers. Given that *LFY*'s effect is through varying the levels of the protein and not just the presence or absence of the protein, increasing *LFY*'s binding affinity for DNA may have led to the transcription of novel proteins or the ability to vary the transcription of previously transcribed genes, thereby making variation in *LFY*'s coding sequence instrumental in the appearance of angiosperms on earth (Hamès et al 2008).

The regulation of flowering time and up-regulation of *LFY* is accomplished through activation of one of four different mechanisms as previously seen in Figure 6. The four pathways consist of a photoperiod related pathway that is induced by genes such as *CRYPTOCHROME 2* (*CRY2* blue light receptor) and *PHYTOCHROME A* (*PHYA* far-red receptor). These activate the *GIGANTA* (*GI*) to *CONSTANS* (*CO*) to *FLOWERING*



*LOCUS T (FT)* to *AGL24/ SHORT VEGETATIVE PHASE (SVP)/ SUPPRESSOR OF OVEREXPRESSION OF CONSTANS (SOC1)* pathway ending in up-regulation of *LFY* (Gregis et al 2008). The photoperiod pathway is repressed the red light receptors, *PHY B*, *D*, and *E*. The vernalization pathway is promoted by genes like *VERNALIZATION 1 (VRN1)*, *VRN2*, and *VIN3* and is prevented by *FRIGIDA (FRI)* through the up-regulation of *FLOWERING LOCUS C (FLC)* (Kuittinen et al 2004). Different *A. thaliana* species have shown that exposure to temperatures of 4°C or less for forty continuous days will fulfill the vernalization requirement (Sung and Amasino 2005).

The primary repressor of flowering is *FLC*, a MADS-box transcription factor. *FLC* is responsible for repressing *SOC1*, *FT*, *FRUTFULL (FUL or AGL8)*, and *LFY* (Edwards et al 2006). *LFY* is inhibited temporally as well by *TERMINAL FLOWER1 (TFL1)*, a phosphatidylethanolamine binding protein (PEBP) (Ohshima, S., et al 1997). *TFL1* represses *LFY* in shoots and they switch roles with *TFL1* being repressed in flowers (Parcy, Bomblies, and Weigel 2002) (Schultz and Haughn 1991).

The other two pathways are the gibberellin (GA) associated (short days) and the autonomous pathways (age-related). The GA pathway affects *LFY* directly through its GA-MYB response element (Eriksson et al 2006). In relation to the autonomous mediated pathway, the genes *FCA* and *LD (LUMINIDEPENDENS)* are two of the primary genes responsible for down-regulating *FLC* and releasing the block on the up-regulation of *LFY* transcription (Samach et al 2000). Mutants of these two genes will

flower sooner than wild type plants if exposed to vernalization (Wang et al 2007). Since LFY is expressed continuously throughout the life of the plant, the Autonomous pathway serves as a backup fail safe mechanism to flowering and is related to the plants age and the overall accumulation of LFY protein.

The multifaceted role of *LFY* in the phase change from vegetative to reproductive has been well documented. *LFY* is known to be transcriptionally regulated by genes of several different types of gene families such as the MADS-box (*API*, *SOC1*, *AGL24*), Homeodomain (*STM*), MYB (*GA-MYB*), AP2-domain, and PEBP (*TFL1*). LFY can also bind and up-regulate itself (Simpson and Dean 2002). The cis-effect of DNA sequence variation in *LFY*'s introns, flanking regions, and UTR's will have the most impact here. Previous studies have shown that most of the variation in *LFY* lies in the introns (Olsen et al 2002).

The LFY transcription factor has many known targets including the homeotic genes *API*, *AP3* and *PI* (with the help of the F-box protein UFO), *CAL* with the assistance of LMI1 (a homeodomain leucine zipper class I (HD-Zip I) element), and *AG* with the help of the homeodomain cofactor WUS (William et al 2004). LFY also directly activates Leucine-rich receptor kinases, signaling proteins, plant specific MYB's of the R2R3 family, Serine-Glycine rich proteins, and a plant specific TUBBY-like DNA binding factor. The *trans*-effect of DNA sequence variation in the coding regions would have the greatest affect on these types of interactions because changing the conformation

of the protein may inhibit its ability to form homodimers or bind to other cofactors and prevent binding and activating downstream targets (Schmid et al 2003).

Coding sequence variation in *FRI* and *FLC* have been shown to be major contributors to flowering time differences in *Arabidopsis* populations. Dominant allele forms of *FRI* keep the plant in a vegetative state enabling it to survive winter by enforcing the vernalization requirement. The importance of coding sequence variation can be seen in that loss-of-function *fri* alleles only need one of two deletions. One of these results in a premature stop codon appearing in the first exon and the other only causes an amino acid substitution. (Gazzani et al 2003). A single insertion in the coding region of *FLC* can result in an allele that delays flowering in a dominant fashion (Michaels and Amasino 1999). Similar species like *Brassica nigra* have shown that indels in the coding region of the *COL1* gene are correlated with flowering time variation (Österberg et al 2002).

The following research sets out to evaluate the DNA sequence variation between different North American and European populations of *A. lyrata* from environments exerting strongly contrasting selective pressures and with contrasting resource allocation phenotypes, and characterize the sequence changes that have taken place in these populations and between *A. lyrata* and *A. thaliana*. Upregulation of *LFY* is directly correlated with the upregulation of the Homeotic genes necessary for floral development. All of the homeotic genes are MADS-domain genes except for *APETALA2* (*AP2*) which

is considered to be a different family of DNA binding protein. These homeotic genes function as developmental switches that are necessary and sufficient to produce the required floral organs (Weigel 1995, Mandel et al 1992). Class A genes consist of *APETALA1 (AP1)* and *AP2* and they produce sepals in whorl 1. Petals are produced in the overlapping domains of the class A and B genes in whorl 2. Class B genes are *APETALA 3 (AP3)* and *PISTILLATA (PI)* The Class C gene is *AG* and it will make stamens with the help of the Class B genes in whorl 3. The carpels in whorl 4 are the responsibility of the Class C gene *AG* by itself and it also represses *AP1* function here (Chae et al 2008) (Jofuku 1994). *LFY* has been shown to work in an activating and repressing manner in these four different whorls. In whorls 1 and 2 *LFY* represses *AG* expression and *LFY* along with the cofactor *WUS* up-regulates *AG* in whorls 3 and 4 (Huala and Sussex 1992).

*AP3* is negatively regulated by *SUPERMAN (SUP)* in whorls 1 and 4 (William et al 2004). *AP3* undergoes positive regulation through the influence of the *LFY/SCF<sup>UFO</sup>* complex. In addition the *SCF<sup>UFO</sup>* complex is also responsible for establishing the dividing line between the meristem cells and organ founder cells being acted on by these developmental proteins (Samach 1999).

*STM* induces the transcription of the F-box protein *UFO*, which is expressed in meristems throughout development. The F-box proteins share a conserved 50 amino acid region. *UFO* expression appears to be involved in setting the boundary between meristem

cells and organ founder cells. The UFO F-box protein interacts with ASK1, another protein that is part of the  $\text{SCF}^{\text{UFO}}$  complex or Skp1/ASK1 (Arabidopsis S-phase Kinase associated Protein1-like1)/Cullin/F-box protein involved in protein degradation via the ubiquitin pathway. UFO may be involved in breaking down stem cell factors and/or cell cycle regulators (Samach 1999)(Chae et al 2008). In *A. thaliana* the *ASK1* gene is also part of the  $\text{SCF}^{\text{TIR1}}$  and  $\text{SCF}^{\text{COI1}}$  F-box protein complexes responsible for initiating the response to the presence of the plant hormones auxin and jasmonate (Kipreos and Pagano 2000).

The UFO F-box protein is the substrate binding portion of the complex as seen in Figure 8 and it binds LFY while the UBC, ubiquitin conjugating enzyme, adds the ubiquitin molecule to LFY targeting it for destruction by the

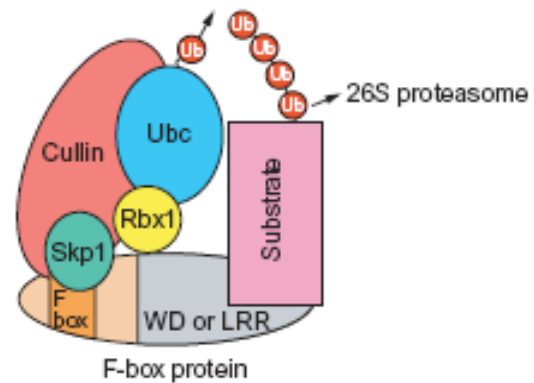


Figure 8 UFO/SCF complex Kipreos, E.T., and Pagano. M. (2000)

26S proteasome. This mechanism may allow for the fine regulation of LFY in a bind/release - bind/release etc. cyclic type manner in the S-phase of the cell cycle versus just the presence or absence of the protein. This could be likened to quickly alternating notes on a flute versus continuously playing the same note or not playing it at all. Conversely the mechanism may also be in place to immediately disassociate LFY after binding. The actual manner in which this action takes

place is not known but variations in the coding region of *LFY* could alter the proteins structure enough to interfere with *UFO*'s binding. Proper binding of *LFY* and *UFO* are necessary for transcription of both of the Class B organ identity genes (Sharma and Fletcher 2002, Kong et al 2004).

In addition to the specific combinations of genes necessary in each whorl, the proteins themselves require proper dimer formation in order to work. *AP1* and *AG* will form homodimers with themselves, *AP1/AP1* and *AG/AG*, and *AP3* and *PI* will form a heterodimer *AP3/PI*. This dimer specificity shows how important alterations in coding sequences can be. Alterations that affect protein conformation can result in disrupting the binding potential of associated proteins and confer altered phenotypes or even bring about the death of the plant. If *AG* is not dimerized correctly for example, whorls 3 and 4 will not form and the plant will not be able to produce any gametes (Reichmann, Krizek, and Meyerowitz 1996).

The activation of the floral induction pathway requires the induction of homeotic genes as previously seen in Figure 6. These genes are responsible for the regulation of body shapes and patterns as well as the initiation and formation of organs. Originally discovered in *Drosophila*, these genes in vertebrates are called *HOX* genes. They contain a conserved sequence known as the homeobox, a region that consists of roughly 180 base pairs and makes a DNA binding domain of approximately 60 amino acids called the homeodomain. They bind to DNA through a helix-turn-helix motif. The binding site

motif is found in two forms CAAT(A/T)ATTG (BS-#1) and CAAT(G/C)ATTG (BS-#2) (Laughon 1991). Not all homeobox genes are HOX genes. HOX genes are just a subgroup of the homeobox gene family. In *Arabidopsis* two well known homeodomain containing genes that *LFY* interacts with are *SHOOT MERISTEMLESS (STM)*, of the *KNOX* family, and *WUS*, of the *KNOTTED1* family. However most of the homeotic genes in *Arabidopsis* are members of the MADS-box gene family and *LFY* can be induced or repressed by them. The term MADS-box domain actually makes reference to the genes in which it was first found. The M is from *MCM1* in *Saccharomyces cerevisiae*, A for *AGAMOUS* in *A. thaliana*, D for *DEFICIENS* in *Antirrhinum majus*, and *SRF* in humans. The MADS-boxes are typically 168 to 180 base pairs long or roughly 56 to 60 amino acids. The protein products of these genes bind to a conserved DNA sequence that consists of a CC(A/T)<sub>6</sub>GG motif (Melzer, Verelst, and Theißen 2009). Mutations in the coding sequence of *LFY* may alter its ability to bind and activate the genes. Without proper activation of these homeotic genes, *Arabidopsis* is unable to initiate the floral pathway and give rise to flowers (Weigel 1995). The homeodomain and homeotic genes are so important for correct plant development that their binding motifs in our *LFY* sequences will be included for analysis. Given the vast differences between the animal and plant homeotic genes they must have evolved along different lines (Bürglin 1996, Hamès et al 2008). The inflorescence meristems are the transition state and will produce floral meristems that will be responsible for producing the sepals, petals, stamens, and

carpels (Sharma and Fletcher 2002). These organs develop in concentric rings or whorls unless there is a mutation in one or more of the homeotic genes controlling their

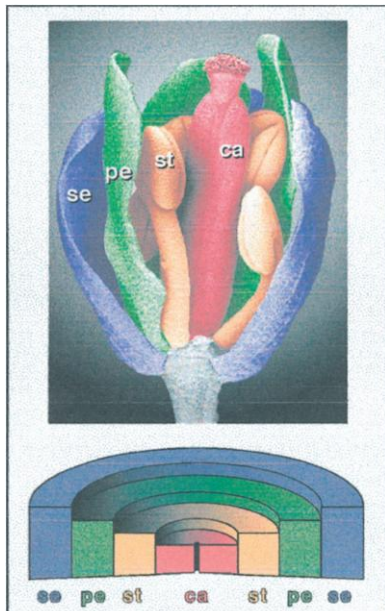


Figure 9 SEM colorized images of a WT *Arabidopsis* flower. Benfey and Weigel 2001

development (Weigel 1995). Mutations can result in the reversion of organ fates or their complete substitution with another organ. The continual observation of these mutant phenotypes led to the concept of the ABC model for flower progression (Benfey and Weigel 2001). In color-coded Figure 9, the top view is a scanning electron microscope image of a nearly mature *Arabidopsis* flower, with sepals (se) first, petals (pe) second, stamens (st) third, and carpels (ca) fourth. The middle image shows the arrangement of the four whorls with

sepals (se) being the first to develop in whorl 1 and carpels (ca) last in whorl 4. The bottom image shows the ABC model, identifying the domains of A, B, and C homeotic activities in a wild type flower. In this model the A domain function alone specifies sepals; domains A plus B petals; B plus C stamens; and C alone carpels. The MADS box gene *AP1* is expressed in the A domain, the MADS box gene *AP3* and *PI* in the B domain, and the MADS box gene *AG* in the C domain. (Benfey and Weigel 2001).



The combinatorial role of the Class A, B, and C genes is seen in the concerted way they work together to form the floral organs. Phenotypes of several homeotic mutants indicate that they alter floral organ identities. The first one of these was the homozygous *ag* mutants (absent gametes) that produced double flowers and no stamens or carpels as seen in Figure 10 a WT flower versus Figure 11 a homozygous *ag* mutant (Ma, Yanofsky, and Meyerowitz 1991).



Figure 10 WT flower *A. thaliana* Page and Grossniklaus 2002 doi:10.1038/nrg730



Figure 11 homozygous *ag* mutant *A. thaliana* Page and Grossniklaus 2002 doi:10.1038/nrg730

The effect of losing one of the floral organ identity genes, like the Class A gene *AP2*, can be seen in Figure 12 (Jofuku et al 1994). In Figure 12 pictures A and B of the WT-flower show the presence of four sepals, two medial (S-M) and two lateral (S-L) in the first whorl. The second whorl has four

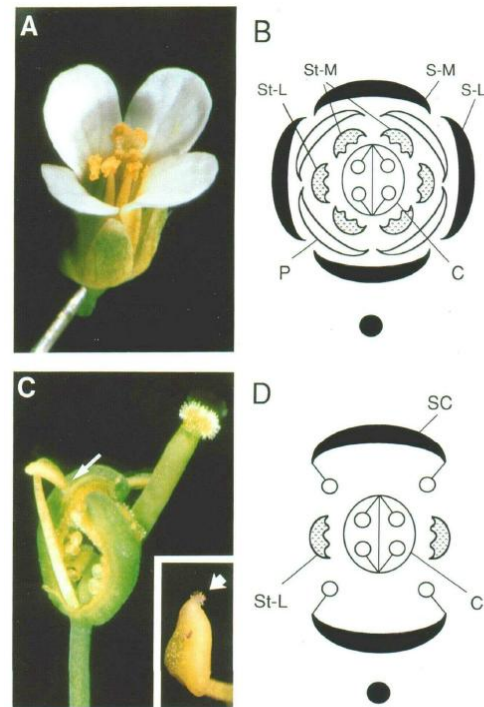


Figure 12 WT flower versus *ap2-10* mutant Jofuku et al 1994

petals (P) that alternate with the sepals. The third whorl has six stamens, four long medial (St-M) and two short lateral (St-L). The fourth whorl has two fused ovule-bearing carpels (C). The *ap2-10* mutant is characterized by the presence of sepal-carpels (SC) in whorl 1, missing 2<sup>nd</sup> whorl petals and 3<sup>rd</sup> whorl medial stamens (St-M), has deformed lateral stamens (St-L) in the 3<sup>rd</sup> whorl, and two normal fused carpels in the 4<sup>th</sup> whorl. The filled circle shows the location of the inflorescence stem.

The conversion from a vegetative stage to a reproductive stage is initially seen as bolting or stem elongation, and may be related to the activity of *STM*. This mechanism is specific to rosette plants such as *Arabidopsis*. Bolting is followed by the formation of cauline leaves with axillary buds and then flower initiation (Huala and Sussex 1992).

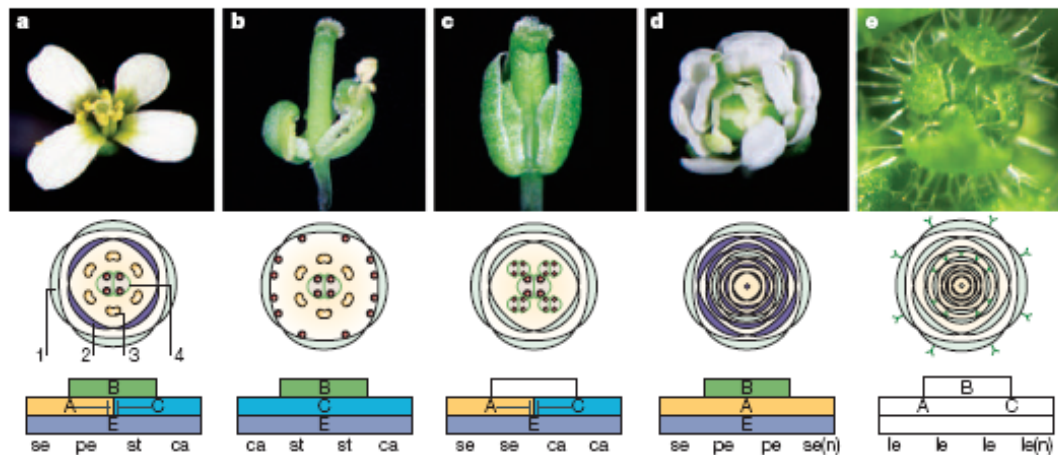


Figure 13 The ABCDE Model mutant phenotypes: a=WT, b=*ap2*, c=*pi*, d=*ag*, e=*sep* Krizek and Fletcher 2005

The ABC model has continued to evolve over the years and now has been expanded from the ABCE model in Figure 13 to the ABCDE model which includes gene specific to ovule identity. In *Arabidopsis* the class E genes are also MADS-box genes and were previously considered to be called *AGAMOUSLIKE2* (*AGL2*) genes and now have been designated *SEPALLATA*

(*SEP*) - *SEP1*, *SEP2*, *SEP3*, and *SEP4* (Melzer, Verelst, and Theißen 2009). *SEP* proteins are unique in the fact that they are the only components required for

formation of all the ABCE

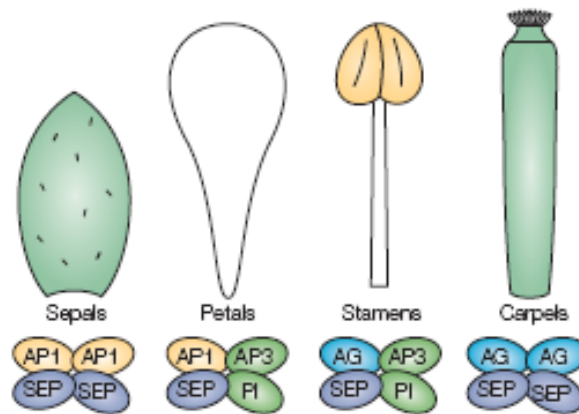


Figure 14 The Quartet Model of MADS-box complexes Krizek and Fletcher 2005

quartet-complexes responsible for floral organ formation as shown in Figure 14. The importance of *SEP* is seen in the partial redundancy of function between the *SEP* proteins. Only triple or quadruple *sep* mutants result in the formation of leaf-like structures instead of the floral organs. *SEP* does not show any effect on the expression of the floral organ identity genes just the actual formation of the quartet-complex (Kater, Dreni, and Colombo 2006).

The Class D genes, responsible for proper development of the ovule including fruit dehiscence, were originally discovered in the *Petunia* and designated *FLORAL*

*BINDING PROTEINS 7 and 11 (FBP7 and FBP11). In Arabidopsis, three members of the AGAMOUS-LIKE (AGL) MADS-box family with overlapping regions of expression were found to occupy Class D, they were identified as SEEDSTICK (STK or AGL11), SHATTERPROOF1 (SHP1 or AGL1) and 2 (SHP2 or AGL5). Experiments have shown that all four members of the AG clade, AG, STK, SHP1, and SHP2 are required for proper formation and function of the ovule (Krizek and Fletcher 2005, Battaglia et al 2006).*

The following research sets out to evaluate possible mechanisms for the phenotypic variation seen in *A. lyrata* populations by isolating and characterizing coding and non-coding sequence differences in *LFY*.

## CHAPTER II

### METHODS

#### *General Methods*

##### Sample Population Information

The source of the *A. lyrata* genome sequence was from a population in Indiana Dunes Michigan, USA. This population is from a cool climate and has an unknown allocation strategy. The field samples came from four different populations, two from Europe and two from North America (NA). One of the North American populations is from Mayodan NC, USA and its location is considered to be on the southern end of the North American *A. lyrata* range. These individuals reside in a warm continental temperate climate and have a heavy allocation of resources (reproductive -versus- vegetative) towards current reproduction with a large number of flowering shoots that have a large number of flowers per shoot. The other NA population is from Ithaca NY, USA. Ithaca has a moderate continental climate, with cold, snowy winters and hot, humid summers. Individuals growing in this region are south of the middle of the entire growing range. Our European populations include individuals from Plech Germany. These individuals are in a cool maritime-influenced climate located north of the middle of the *A.*

*lyrata* range and have a medium resource allocation strategy. The fourth population is from Spiterstulen Norway and lives in a boreal maritime-influenced climate on the northern end of the *A. lyrata* range. Its resource allocation strategy is considered to be heavy towards vegetative development (Remington lab unpublished data 2009). For comparison I also included *Boechnera drummondii*, a more distant relative of *A. lyrata* and a member of the Brassicaceae family. The available data was limited and a complete comparison across the entire region we sequenced was not possible.

### **PCR Amplification, Cloning, and Sequencing:**

The *A. lyrata LFY* standard sequence was determined by using NCBI BLAST software on 69 *A. lyrata* trace archives using *Arabidopsis thaliana* query sequences and assembled contigs. DNA samples of field samples were collected from fresh and frozen plant tissue using the NucleoSpin Plant Kit by Macherey-Nagel. PCR and Sequencing primers were designed with PRIMER3 using the *A. thaliana* and *A. lyrata* standards. The terms “U” and “L” denote upper 5’ to 3’ and lower 3’ to 5’ directions respectively (Table 2.1). Vector primers M13F and M13R were supplied by MWG/Operon. The entire *LFY* region was amplified with the primers *LFY* 1161U and *LFY* 4908L. Sequencing used those two primers in addition to Int1bU, 2<sup>nd</sup> Lower, Int2U, 3<sup>rd</sup> Lower, Primer C-U, and 5<sup>th</sup> Lower (Table 2.1).

**Table 2.1 PCR and Sequencing Primers**

<b>Primer</b>	<b>Sequence</b>	<b>Region</b>	<b>Purpose</b>
Int1bU	cttgatgctctctccaagaa	Intron 1	cloning and sequencing
Int1bL	gtctggtttgctgttgac	Intron 1	initial PCR
Int2U	gccgtgagttccttctcag	Intron 2	cloning and sequencing
Int2L	cgcattttgggcttgttat	Intron 2	cloning and sequencing
Primer A-U	aaaaatgcggaggatgaaaa	5' UTR	initial PCR
Primer A-L	tcagataaccctgtccaatca	begin exon 2	initial PCR
Primer B-U	tgcaagaagtacgaggattca	end intron 1	cloning and sequencing
Primer B-L	caactaactacacccaacgaaaa	middle intron 2	initial PCR
Primer C-U	cggcggataatagagggtct	middle intron 2	cloning and sequencing
Primer C-L	caacctagatgaccatatgtttga	3' flanking	initial PCR
Lfy 1161U	cgtgctctcatgatgcaaac	5' UTR	cloning and sequencing
Lfy 4908L	caccagtaaatcggtttcg	3' flanking	cloning and sequencing
2 <sup>nd</sup> Lower	tgtgtatggcatcaaaacaat	middle intron 1	cloning and sequencing
3 <sup>rd</sup> Lower	gttccctaccataccat	begin intron 2	cloning and sequencing
5 <sup>th</sup> Lower	gacgacaagcaatgttcac	middle exon 3	cloning and sequencing

PCR fragments were generated with a QIAGEN LR-PCR kit (cat #1043031) using the manufacturers PCR cycling protocols in the QIAGEN LongRange PCR Handbook and visualized with gel electrophoresis. Fragments were subjected to Nanodrop spectrophotometer analysis to determine nanogram per microliter concentration. The volume of the PCR samples was reduced using the Speed-Vac to achieve a 50 ng/μl concentration. PCR samples were labeled with the DYEnamic ET Terminator Kit (Amersham Biosciences cat #USB 1090/1095) for the MegaBACE 1000

sequencer by thermally cycling them one time with a reaction pre-mix and our primers according to the manufacturers' protocols (Amersham Biosciences). Reaction products were precipitated and concentrated to remove unincorporated dye-labeled terminators and then resuspended in a MegaBACE loading solution (Amersham Biosciences), separated, and detected on the MegaBACE 1000 in Dr. V. Henrich's lab. The samples were run at three different voltage settings (voltage times time) in order to obtain sufficient signal strength without overloading samples and generating poor sequence quality due to slow runs.

The recovered sequences were initially aligned by Bioedit/BioLign software (<http://bioedit.software.informer.com/>) and then manually by sight. They were then subjected to PHRED/PHRAP software (David Remington lab) analysis and the few high quality sequence regions recovered were aligned. The six individuals that provided the preliminary data were Mayodan M0611-10, M0634-1, M06030-1 and Plech P3-7, P8-37, and P4-5. Preliminary functional genetics analysis was performed by retrieving known transcription factor binding site motifs from the Database for Arabidopsis Transcription Factors website at <http://datf.cbi.pku.edu.cn/>. The presence or absence of known motifs was confirmed in Bioedit software by utilizing the user-defined motif search function with wildcard characters for motif positions with multiple nucleotide possibilities.

We successfully amplified the entire *LFY* region with the primers *LFY* 1161U and *Lfy* 4908L from 13 individuals (the # indicates how many clones were recovered): I1-1 (1), I1-1B (1), N1-3 (3), N3-9 (1), N9-1 (4), N10-14 (4), P3-1 (4), P6-12 (4), S1-1 (4),



S5-21A (2), S5-21B (2), S9-4 (4), AND S10-4 (4). We used an Invitrogen TOPO TA Cloning Kit for Sequencing (cat #K475-40) to make the clones. Insert DNA was extracted with a QIAprep Spin Miniprep Kit (50) (cat # 27104). Insert DNA was released from vector by EcoRI restriction enzyme digestion and insert size confirmed by DNA gel electrophoresis. Plasmid DNA was quantified using a Nanodrop spectrophotometer and the volumes were adjusted. DNA sequencing was performed by Eurofins MWG-Operon and sample data was retrieved via their website <http://www.operon.com/default.aspx>.

### **DNA Sequence Analysis:**

The .ace files were initially aligned by BioLign software and then manually by sight in BioEdit to create our consensus sequences. If samples produced more than one allele they were grouped together and given designations such as “AD”, “BC”, etc. If a sample only produced one allele then they were not given any extra designation. Samples that were incomplete due to primer failure or did not meet our minimum base call score of PHRED 40 from the PHRED/Phrap software analysis were not used. Base calls with quality scores less than PHRED 40 are considered unreliable. Sequences that met our quality score specifications and contained the entire *LFY* region amplified were identified as follows: I1-1, N1-3, N9-1, N10-14, P6-12AD, P6-12BC, S1-1AC, S1-1BD, S5-21A, S5-21B, S10-4ABD, and S10-4C.

The qualified sequences were assembled in BioEdit and visually characterized for the location of the known structural elements, conserved domains, indels, and polymorphisms. The presence or absence of known motifs was confirmed in Bioedit

software by utilizing the user-defined motif search function with wildcard characters for motif positions with multiple nucleotide possibilities.

Sequence inserts were analyzed using NCBI BLAST software for known identities. Sequences associated with miRNAs were subjected to further analysis. For the miRNA analysis, candidate sequence inserts were copied and pasted in fasta format into the search box in the “DNAfold” portion of the “mfold” website on the Rensselaer bioinformatics web server at <http://mfold.bioinfo.rpi.edu/cgi-bin/dna-form1.cgi> in order to determine if they would assume a typical stem loop structure. The default settings were used. Pre-qualified sequences were copied and pasted directly into the single sequence search box with the “Search Sequences option” set at “Stem-loop sequences” on the miRBASE website at <http://www.mirbase.org/search>. The associated miRNA was listed with its core sequence and any associated miRNA family members.

The population genetics analysis was performed with DnaSP software from <http://www.ub.edu/dnasp/> website. The previously analyzed sequences were imported into DnaSP and identified by categories: genome standards (*A. lyrata*, *A. thaliana*, *B. drummondii*), all *A. lyrata*, *A. lyrata* NA, and *A. lyrata* Europe. The individual sequences were divided into coding and non-coding regions. The test performed were the MacDonald/Kreitman test, Polymorphism and Divergence, Tajima’s D, and Fay and Wu’s H.

## CHAPTER III

### RESULTS

The BioEdit alignment of 69 *A. lyrata LFY* contigs from the *A. lyrata* genome sequencing project (<http://www.phytozome.net/alrata.php>) gathered from NCBI's online database generated the consensus sequence used as the *A. lyrata LFY* standard for this project. Compared to the *A. lyrata LFY* standard, a second *A. lyrata LFY* sequence in the public databases (Olsen et. al. 2002) had an 88 base pair (bp) insertion, consisting of a 22 bp insert repeated four times, in intron 1. This repeat leads directly into exon 2 and contains the first 15 nucleotides (NTs) of exon 2. The 88 bp insert was not found in *A. thaliana*.

Based on these findings we designed primers for intron 1 and amplified this region from six individuals from Spiterstulen, Norway. Our results showed the presence of multiple sequence length variations for intron 1. A multiple restriction enzyme digest assay also revealed the presence of multiple fragment length variants and heterozygous individuals (data not shown).

The 2<sup>nd</sup> intron of *LFY* was amplified from three new Spiterstulen samples (SX-30, SX-33, SX-34), one sample from Plech, Germany (P3-7), and two samples from Mayodan, North Carolina, USA (M611-10, M634-5). The primer combination amplified

all samples and showed the presence of multiple fragment length variants and heterozygous individuals as shown in figure 15. Introns 1 and 2 produced fragments with four different lengths. The Spiterstulen samples contained heterozygous individuals and the longest and shortest fragment lengths. The Plech sample was homozygous and resembled the shortest Spiterstulen fragment. The North Carolina samples were also homozygous and had fragments approximately the same length as the longest Spiterstulen samples.

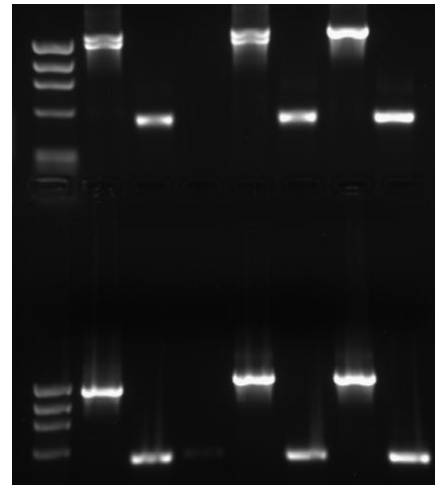


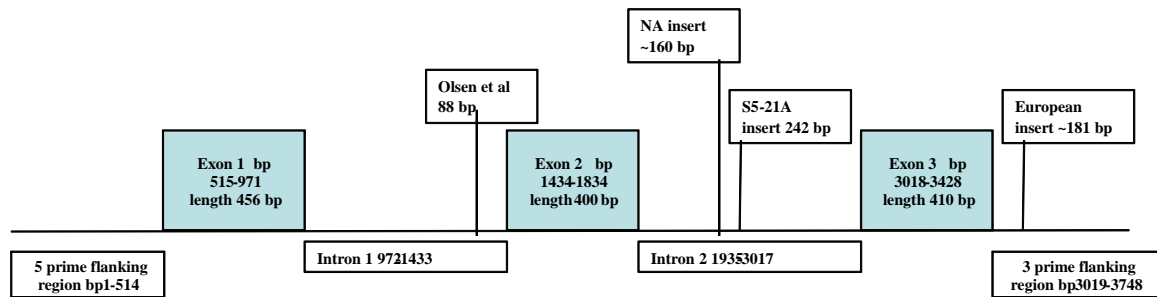
Figure 15 PCR results from amplification of *LFY* intron 1 and 2. Top row: Lane 1 Phi-x ladder, Spiterstulen samples: SX-30 int2, SX-30 int1, blank, SX-33 int2, SX-33 int1, SX-34 int2, and SX-34 int1. Bottom row: Phi-x ladder, Plech samples: P3-7 int2, P3-7 int1, blank, M611-10 int2, M611-10 int1, M634-5 int2, and M634-5 int1.

Three new overlapping primer combinations were designed to amplify the entire *LFY* gene region including all three exons, some of the 5' and 3' prime flanking regions and both of the introns. We attempted multiple PCR reactions on each of the three primer combinations with varying levels of success but eventually amplified enough samples to sequence. The results of the three different MegaBACE sequencing runs were poor. More than half of each of the recovered fragments was unusable due to indel heterozygosity; however, performing a NCBI BLAST software search on the good quality reads confirmed the fragments were from *A. lyrata* and these were easily aligned to the *A. lyrata* standard. Analysis also showed the presence of predicted conserved transcription factor binding site motifs amongst all six *A. lyrata* individuals in figure 15 including the *A. lyrata* and *A. thaliana* reference sequences. The only differences detected in the

binding site motifs were the presence or absence of individual motifs between the two species *A. lyrata* and *A. thaliana*.

Multiple attempts to amplify a single fragment containing the entire *LFY* gene region including all three exons, some of the 5' and 3' prime flanking regions and both of the introns had mixed results. A primer combination covering a 3748 bp fragment (figure 16) using one individual from each of four populations (Mayodan, Plech, Spiterstulen, and Ithaca, New York) was successful. This combination recovered fragments of varying lengths along with homo- and heterozygous individuals.

Sixteen individuals, four from each population, were amplified by PCR using the 3748 bp primer combination and produced mixed results. PCR products from successful reactions were cloned and transformed to improve our sequence quality. We were able to obtain 38 clones which produced 15 alleles for sequencing. The overall quality of the recovered sequence reads was good except for the region intron 1 amplified by the sequencing primer 2<sup>nd</sup> lower. That primer failed to amplify more than half of the time due to the presence of multiple indels.



**Figure 16 – *LEAFY* schematic showing locations of major inserts**

Table 3.1 – Sequenced *LFY* alleles and number of clones by population and region.

Population	Allele	# Clones
<i>N. America:</i>		
Mayodan	N1-3	3
	N9-1	4
	N10-14	4
Ithaca	I1-1	1
<i>Europe:</i>		
Plech	P3-1AC	2
	P3-1BD	2
	P6-12BC	2
	P6-12AD	2
Spiterstulen	S1-1AC	2
	S1-1BD	2
	S5-21A	2

Population	Allele	# Clones
	S5-21B	2
	S9-4	4
	S10-4ABD	3
	S10-4C	1

### **Final Sequence Data Analysis:**

All sequences were aligned, and compared using the *A. lyrata* standard sequence as the reference sequence unless stated otherwise. Sequence position locations (+/- bp) were based on the alignment of each sequence and numbered relative to the first position of the start codon in the *A. lyrata* standard sequence used in this analysis.

### ***A. lyrata* genome standard versus *A. thaliana* genome sequence:**

*A. thaliana* had a 36 bp insert and a 19 bp deletion upstream of the start codon. *A. thaliana* had 18 substitutions in exon 1, 16 substitutions and a 6 bp deletion in exon 2, and 18 substitutions in exon 3. In intron 1 *A. thaliana* had an 11 bp deletion and an 18 bp insertion. Intron 2 of *A. thaliana* had multiple indels of varying lengths including insertions of 16 bp, 96 bp, and 16 bp and deletions of 32 bp, 14 bp, 57 bp, 11 bp, 68 bp, 12 bp, 15 bp, and 12 bp. Intron 2 and the 3' flanking region were the hardest to align due to multiple indels, including many not listed between 1 and 10 bps.

### ***A. lyrata* genome standard vs. other *A. lyrata* alleles:**

The sequence identified as Olsen et. al. 2002 has an unknown origin and was only analyzed in the context of total *A. lyrata* group comparisons to *A. thaliana* (Olsen et. al.

2002). In the 5' flanking region all of the samples recovered had deletions ranging from 6-21 bp in the "AG" repeat upstream of the start codon. The only exception to this was the Mayodan sample, N1-3, that matched the *A. lyrata* standard exactly across the "AG" repeat. The other North American (NA) samples had deletions of: N9-1 (-12 bp), N10-14 (-6 bp), and I1-1 (-10 bp). The European samples had only two length variations of -16 and -21 bp in this region. All four of the Plech samples and five of the Spiterstulen samples had the exact same 21 bp deletion. The other two Spiterstulen samples, S10-4ABD and S10-4C, each had the same 16 bp deletion but retained five bp the other European populations did not have. The rest of the 5' flanking region had a few SNPs that were either population specific or geographically localized.

Exon 1 was highly conserved with all of the *A. lyrata* samples showing a T to C polymorphism. The European alleles P6-12, P3-1, S1-1, S5-21, and S9-4 have an A to G polymorphism.

Intron 1 was highly conserved with the majority of polymorphisms ranging from 1-8 nucleotides occurring in the European populations. In the NA populations, Ithaca had a 10 bp insert of five "TA" repeats. The European populations are highlighted by a 56 bp insert.

Exon 2 was highly conserved with the only polymorphisms occurring in the European populations. As a group, the alleles P3-1BD and S1-1AC/BD have an A to G polymorphism, G to T, and a T to C. Also as a group the alleles P3-1AC/BD, S1-1AC/BD, and S10-4ABD/C have C to T polymorphism.



Intron 2 had the highest level of diversity and contained the longest insertions. The populations from NA had an almost identical insert ranging from 158 bp to 161 bp. The only differences found between these sequences were that N9-1 had a 3 bp deletion of As and I1-1 had two individual T deletions along with a TT insert and a G to A substitution. Amongst all populations, the European sample S5-21B had the largest insert of 242 bp. S5-21B also has a 39 bp insertion at the same location as the NA populations.

Exon 3 had the highest level of diversity among the coding regions with the majority of variation again located in the European populations. Variation ranged from 3bp deletions to 12 bp insertions and included as many as 5 polymorphisms. The NA populations only contained two polymorphisms total. The European alleles P6-12BC and P6-12AD contain a 12 bp insert and alleles S5-21B, S9-4, and S10-4C have a 3 bp deletion.

The 3' flanking region was highly conserved with the exception of two inserts found only in European populations. Alleles S5-21B, S9-4, and S10-4C have a 38-41 bp insert. The only difference is the presence/absence of two or three T's in a run of 19 T's. All of the European alleles except P6-12BC and P6-12AC have an 83 bp insert. Allele S10-ABD is missing three nucleotides from this insert and it has a length of 80 bp. S10-ABD also has a 12 bp insert. The total length of the European populations' recovered fragments varied from 3786 bp to 4095 bp, a difference of 309 bp. The NA populations had the least variation in length, 3893 bp to 3908 bp, a total of only 15 bp. (See Appendix A.1 for the lengths of the regions sequenced and number of polymorphism present as compared to the *A. lyrata* standard)

### mfold / miRBASE Sequence Analysis

Many transposable elements have inverted terminal repeats, and miRNAs form stem-loop structures. Thus, our sequence inserts (Fig. 16) were subjected to mfold analysis to confirm the ability to assume an RNA stem loop structure. Three candidate structures were found, two in the Europeans and one in NA. The largest insert of 241 bp was found in intron 2 of the Spiterstulen sequence S5-21A and this fragment is found on all five chromosomes of *A. thaliana*. The fragment was identified as a little-Athila Ty3/Gypsy LTR retrotransposon. Based on the *A. thaliana* genome sequence this fragment is found on

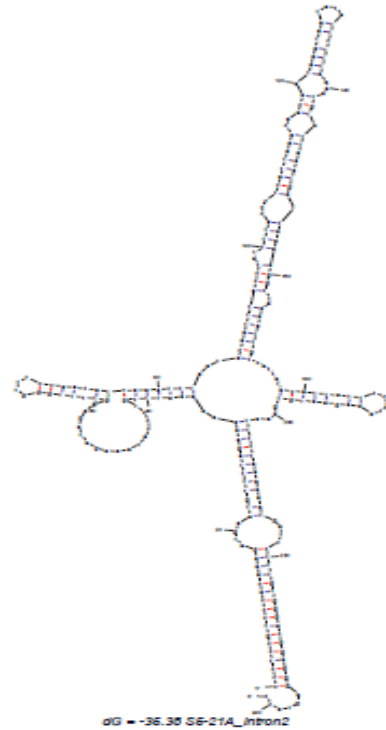


Figure 17 Intron 2 insert S5-21A folded by "mfold" software.  
<http://mfold.bioinfo.rpi.edu/cgi-bin/dna-form1.cgi>

Chromosome 5 but not in the *LFY* gene region. The Mfold website produced the image in Figure 17.

The second largest fragment, ~161 bp, is also located in intron 2 and is only found in all of the NA populations. This fragment has been identified in *Brassica rapa* "Chinese cabbage" and was identified through a search of the miRBASE database as an miRNA ath-MIR157. The ath-MIR 156/157 family is found on all 5 *A. thaliana* chromosomes but not in the *LFY* region. The Mfold website produced the image in Figure 18.

The last fragment analyzed is from the European alleles P3-1, S1-1, S5-21, S9-4, and S10-4. This fragment is from the 3' flanking region and was identified as ath-MIR169 found on *A. thaliana*'s 3<sup>rd</sup> chromosome. This miRNA is known to target mRNAs coding for CCAAT binding factor (CBF) HAP2-like proteins. The MIR169 family is found on all five of the *A. thaliana* chromosomes. The Mfold website produced the image in Figure 19.

### Polymorphism and Divergence:

The polymorphism and divergence data from the *A. lyrata* versus *A. thaliana* comparison shows the ratio of nonsynonymous polymorphisms 0.274 was 2.36 times higher than the divergence ratio 0.116, The European populations show a much higher rate of polymorphism than the NA populations (Table 3.4).

### MacDonald/Kreitman (MK) Tests:

The MacDonald-Kreitman tests of neutral evolution can be used to infer the ratio of DNA substitutions in coding regions under positive selection.

The test is performed on two different species and is

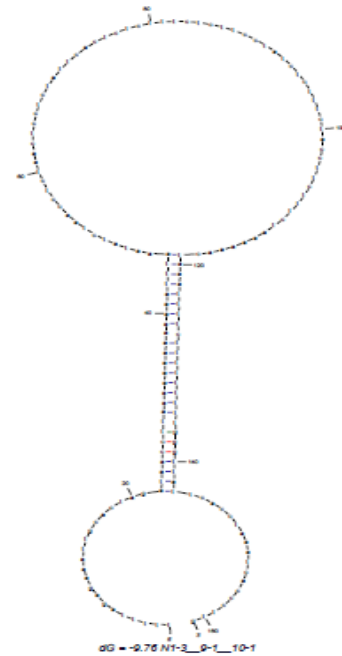


Figure 18 NA insert in intron2 folded by "mfold" software.  
<http://mfold.bioinfo.rpi.edu/cgi-bin/dna-form1.cgi>

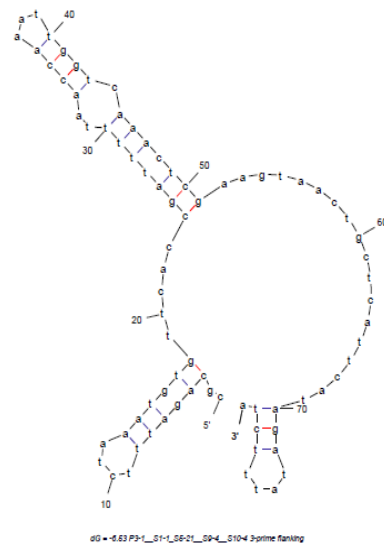


Figure 19 European inserts 3' flanking by "mfold" software.  
<http://mfold.bioinfo.rpi.edu/cgi-bin/dna-form1.cgi>

based on the per-site ratio of non-synonymous to synonymous polymorphisms within species versus the per-site ratio of non-synonymous to synonymous substitutions between species. We used *A. thaliana* as the outgroup for the comparison. The biological relevance is that a higher ratio between species versus within species is considered evidence for divergent selection consistent with adaptive amino acid substitution. A higher ratio within species versus between species is considered evidence for balancing selection or may indicate retention of deleterious amino acid substitutions (Bustamante et al 2002). Under the neutral theory of evolution these ratios should be equal.

**Table 3.4 Summary of polymorphism and divergence rates**

	<i>A. lyrata</i> (pi)	Europe <i>A. lyrata</i> (pi)	NA <i>A. lyrata</i> (pi)	<i>A. lyrata</i> / <i>A. thaliana</i> divergence (K)
Dn	0.00326	0.00384	0.00083	0.01630
Ds	0.01188	0.01262	0.00134	0.14112
Dn/Ds ratio	0.274	0.304	0.620	0.116
Note all values calculated with Jukes Cantor model				

The proportion to which the ratios depart from equality is quantified as the Neutrality Index or NI. If we assume that synonymous sites are evolving neutrally, a NI value of less than one implies an excess of amino acid divergence, on the other hand a NI value of more than one would imply an excess of amino acid polymorphism. I used the test to try and identify if the differences within *A. lyrata* and between *A. lyrata* versus *A. thaliana* were contributed to divergent or balancing selection. The MK tests showed that *A. lyrata* had significantly elevated levels of non-synonymous polymorphisms, whether *B. drummondii* or *A. thaliana* was used for the interspecific comparison (Tables 3.1 and 3.2). The European *A. lyrata* alleles had significantly elevated levels of non-

synonymous polymorphisms compared to *A. thaliana* (Table 3.3). The 2x2 contingency table could not be calculated for North American samples due to the low number of available NA samples.

**Table 3.1 *A. lyrata* versus *B. drummondii* McDonald-Kreitman test results using combined *A. lyrata* samples**

NI: 6.325 with a Fisher's exact test P-value (two tailed) of 0.000075*** (***) P<0.001).	Interspecific substitutions	<i>A. lyrata</i> polymorphisms
Synonymous	44	16
Non-synonymous	10	23

**Table 3.2 *A. lyrata* versus *A. thaliana* McDonald-Kreitman test results using combined *A. lyrata* samples**

NI: 3.801 with a Fisher's exact test P-value (two tailed) of 0.004727** (P<0.01).	Interspecific substitutions	<i>A. lyrata</i> polymorphisms
Synonymous	35	17
Non-synonymous	13	24

**Table 3.3 Europe versus *A. thaliana* McDonald-Kreitman test results using European *A. lyrata* samples**

NI: 3.834 Fisher's exact test P-value (two tailed) of 0.008964** (** 0.001<P<0.01).	Interspecific substitutions	European polymorphisms
Synonymous	36	13
Non-synonymous	13	18

### **Tajima's D Tests:**

The Tajima's D test is used to determine whether or not the sequences being compared are under the influence of selection or genetic drift. The test compares the allele frequency spectrum of polymorphic sites with what would be expected under neutrality. The test uses  $P_i$ , the average pair wise difference between sites, versus  $\theta$ , the total frequency of polymorphic sites adjusted for the number of sequences according to the neutral coalescent model. The value for D was calculated as a standardized value of

Pi minus theta. The Tajima's D test produced an overall value of -1.67 for all sites. The value was significant for the coding region and nonsynonymous sites. The D value for nonsynonymous sites was substantially more negative than the value for synonymous sites, implying that nonsynonymous polymorphisms are skewed towards rare alleles (Tables 3.5 – 3.7). The same trends were seen when the NA and European samples were analyzed separately.

**Table 3.5 Tajima's D test for *A. lyrata***

Tajima's D: -1.66733
Coding region: Tajima's D: -1.87831    *, $P < 0.05$
Synonymous sites: Tajima's D(syn): -1.17980    Not significant, $P > 0.10$
NonSynonymous sites: Tajima's D(nonsyn): -2.25129    **, $P < 0.01$
Tajima's D (nonsyn/syn) ratio: 1.90819

**Table 3.6 Tajima's D test for *A. lyrata* North America**

Tajima's D: -0.69948
Coding region: Tajima's D: -1.04849    Not significant, $P > 0.10$
Synonymous sites: Tajima's D(syn): -0.81650    Not significant, $P > 0.10$
NonSynonymous sites: Tajima's D(nonsyn): -0.97256    Not significant, $P > 0.10$
Tajima's D (nonsyn/syn) ratio: 1.19114

**Table 3.7 Tajima's D test for *A. lyrata* Europe**

Tajima's D: -1.07754    .
Coding region: Tajima's D: -1.39620    Not significant, $P > 0.10$
Synonymous sites: Tajima's D(syn): -0.70129    Not significant, $P > 0.10$
NonSynonymous sites: Tajima's D(nonsyn): -1.80969    *, $P < 0.05$
Tajima's D (nonsyn/syn) ratio: 2.58052

#### **Fay and Wu's H test:**

The Fay and Wu's H test compares the frequency of derived alleles at polymorphic sites against neutral expectations. A significantly negative H value signifies a high frequency of derived polymorphisms and provides evidence of positive selection. The Fay and Wu's H test resulted in a high positive value, 8.51. This result shows that derived polymorphisms are skewed towards being rare.

### **Location Characteristics of Specific *A. thaliana*/*A. lyrata* Polymorphisms:**

The coding region of the *LFY* alignments encodes 422 amino acids (aa) and is divided into three areas: exon 1 (aa 1-152), exon 2 (aa 153-286), and exon 3 (aa 287-422). The P6-12BC and P6-12AD alleles are longer due to the presence of a four amino acid insert after amino acid 397 in Exon 3. *LFY* contains two conserved regions, the C-Domain and the N-Domain. The N-Domain is restricted to amino acids L<sub>46</sub> to E<sub>121</sub> of exon 1 while the C-Domain spans the latter half of exon 2 and all but the last 34 amino acids of exon 3. The C-Domain contains nine regions; 2  $\beta$ -pleated sheet areas followed by 7  $\alpha$ -helices. In *A. lyrata* populations, exon 1 had the least amount of variation (pair wise avg. ~1 aa change) and exon 3 (1-3 aa change) had the most. In *A. thaliana*, exons 1 and 2 had the most divergence from *A. lyrata* (6 aa and 7aa respectively) while exon 3 had the least (1 aa). Exon 1 contained an M>T substitution in all alleles except the *A. lyrata* standard. All of the amino acid changes found in exons 2 and 3 of *A. lyrata* were in the European populations with the exception of a radical 1 aa substitution, A>T, in the North American allele N9-1 at position 391 of structural element  $\alpha$ 7, just 1 aa past the end of the C-Domain (see Appendices A.2, A.3, A.4 for coding region alignments and locations of important sites known to be involved in DNA sequence and backbone binding, LFY dimer formation, and phenotypic variation).

### **Polymorphism and substitution at functionally important sites:**

**N-Domain – transcriptional regulation (amino acid (aa) 46-121):** Using the *A. lyrata* standard for comparison, the only differences found in the N-Domain were an A>P substitution in the *A. thaliana* sample at position aa 55 and a radical polymorphism, R>T,

in the Olsen et. al. sample at position aa 112. The substitution in the Olsen et. al. sequence was at a known site involved in intermediate level mutant phenotypes.

**C-Domain – DNA binding affinity (aa 231-390):** There were no differences found in the  $\beta 1$  (aa 238-240) or  $\beta 2$  (aa 254-256) regions. Region  $\alpha 1$  (aa 257-280) contained 2 radical polymorphisms. Sample S1-1BD had an R>H polymorphism at position aa 266 and S1-1AC had a Q>R polymorphism at position aa 273. Position aa 293 in region  $\alpha 2$  (aa 288-298) is known to interact with the DNA backbone when binding and P6-12BC had a radical polymorphism, Y>H, at position aa 294. Two more radical polymorphisms were found between regions  $\alpha 2$  and  $\alpha 3$  (aa 305-320). S10-4C had a Y>H at position aa 302 and S5-21A had an I>T at position aa 303. Position aa 304 is known to be involved in binding to the DNA backbone. Structural element  $\alpha 3$  contained a radical polymorphism at a position known to interact with the DNA backbone, aa 311, in P6-12AD. The position at aa 310 is considered critical for the binding affinity of the LFY protein (Hamès et al 2008, Maizel et al 2005, Weigel et al 1992).

The structural element  $\alpha 4$  (aa 322-336) revealed two more radical polymorphisms in the Plech samples. P6-12BC and P6-12AD had an A>S polymorphism at position aa 331. *A. thaliana* had a D>E substitution at position aa 323 and S10-4ABD had a K>R polymorphism at position aa 333. No differences were found in the structural elements  $\alpha 5$  (aa 339-355) and  $\alpha 6$  (aa 361-367). However, P6-12BC contained a radical polymorphism, D>G, at position aa 360 between the structural elements  $\alpha 6$  and  $\alpha 7$ .

The last structural element,  $\alpha 7$  (aa 378-397), had a radical polymorphism, R>W, in S5-21A. This site is known to be involved in *LFY* dimer formation (Hamès et al 2008,



Maizel et al 2005, Weigel et al 1992). It also contains an A>V polymorphism at position aa 393 in P6-12AD.

**Mutations in the remainder of the coding region:** In exon 2 at position aa 164 is a mutation known to be involved in a strong level mutant phenotype and *A. thaliana* and *B. drummondii* have a N>D radical substitution at position aa 163. S10-4BD has a Q>R radical substitution at position aa 190. A known mutation at aa 189 also results in a strong level mutant phenotype. The samples P6-12BC/AD have a four aa insertion after position aa 397 at the end of the structural element  $\alpha 7$ .

**Putative Transcription Factor (TF) Binding Site Motif analysis:**

The search for and cataloguing of eight predicated TF binding sites using forward and reverse sequences against the total fragment alignments produced the following results. The TF binding site motifs for ABA response elements, AP2/EREBP proteins, and (HD-Zip) Homeodomain-leucine zipper containing proteins were not found in this analysis.

In the search for ARF auxin response element TF sites, *A. thaliana* had four sites, *A. lyrata* had two, and *B. drummondii* had one. One potential site in *A. lyrata* was eliminated due to a G>C substitution.

The presence of a known GA-Myb TF binding site located in *A. thaliana* was used as a reference site. All of the sequences were found to have it at the same alignment site. A search for the consensus Myb-element motif found four sites common to all the sequences. The European sample P3-1AC/BD, S1-1AC/BD, S5-21A/B, S9-4, and S10-

4C have an extra site in their 3' insert. Also P3-1BD, S1-1AC and BD have an additional site due to an A>G substitution. *B. drummondii* has two extra sites due to A>T mutations.

The search for self-regulating LFY TF binding site located three candidate sites in all sequences. N9-1 has an additional site due to polymorphism. *B. drummondii* does not have one of these sites due to an insert. Three more potential binding sites were found only in the *A. lyrata* populations. *A. thaliana* is missing one of the sites due to an inset and *B. drummondii* is missing one due to lack of data.

Using the known region of a SOC1 binding site as reference point, the search for the canonical MADS-box motif C(A/T)<sub>6</sub>G found three motifs in all the sequences. S1-1BD did not have one of the sites due to an A>G substitution. All the sequences had an additional site that the P3-1BD, S1-1AC, and S1-1BD samples did not have due to a T>C substitution. *A. thaliana* and *B. drummondii* have deletions that eliminate this site. *A. thaliana* has two sites unique to itself and a C>A substitution eliminates one of these sites in the *A. lyrata* populations.

The last sets of TF binding sites searched for were for WRKY proteins. The search found 9 sites in all the sequences except for *B. drummondii* which has no sequence for two of them. *B. drummondii* has a site not found in the other sequences due to an insertion and a substitution. The Olsen et. al. 2002 sequence did not have one of the sites due to missing sequence data and *A. thaliana* did not have this site due to several substitutions. *A. thaliana* did have one extra site due to a substitution.

## CHAPTER IV

### DISCUSSION

This research project set out to identify and categorize indels and single-nucleotide polymorphisms in the coding and non-coding regions of *LFY* that may be responsible for the wide variety of phenotypes found in the *A. lyrata* population and contribute to the phenotypic variation between the out crossing perennial *A. lyrata* and its closest known relative the inbreeding annual *A. thaliana*. A semelparous species like *A. thaliana* will reproduce once and then die. An iteroparous species, like *A. lyrata*, will reproduce several times over multiple years and can vary in the degree and pattern to which it allocates resources towards reproduction vs. vegetative growth. This variation in resource allocation is attributable to genetic variation and the phenotypic variation in *A. lyrata* is also geographically based. To thoroughly investigate the genetic variation within *A. lyrata* and between *A. lyrata* and the European based *A. thaliana*, the populations chosen for this study covered the entire *A. lyrata* growing range from the northern to southern extremes and were separated into North American and European subgroups.

We chose *LFY* as our candidate gene given its role as the primary regulator of the transition from a vegetative state to a reproductive state in *Arabidopsis*. It has been

shown that variation in *LFY*, including introns, flanking regions, and its coding regions is responsible for when, where, and how much *LFY* gets expressed. The timing, location, and quantity of *LFY* are correlated with phase changes in *Arabidopsis* and variations in these events are associated with phenotypic variation. Changes found in the coding regions may alter the intrinsic properties of *LFY* and changes here may have significant consequences such as altering binding affinities to DNA or dimerization partners (Blasquez et. al. 1997)

The initial between species DNA sequence analysis of the *A. lyrata* standard versus the *A. thaliana* genome standard showed the presence of multiple distinguishing polymorphisms between *A. lyrata* and *A. thaliana*. Genetic variation responsible for quantitative variation in life history phenotypes is unlikely to involve changes in constitutive function, and thus unlikely to occur in highly conserved domains. The N- and C-Domains are considered to be the most conserved regions across all *LFY* homologues and 7 of the 11 identified *A. thaliana* mutant phenotypes are in these two regions.

### **Indel Variation:**

The general sequence overview analysis of the *A. lyrata* standard against our data and *A. thaliana* showed the presence of multiple indels and polymorphisms all across the regions compared. The *A. lyrata* standard uniquely differs from all of our *A. lyrata* alleles, the *A. thaliana* sequence, and the *B. drummondii* sequence at four additional sites potentially reflecting sequencing errors or genuine singleton polymorphisms that may have arisen recently in ancestors of the genotype sequenced in the standard.

### **Coding Region Divergence:**

The level of diversity between species varied greatly between the coding regions. *A. thaliana* had 13 amino acid substitutions and 2 amino acid deletions total compared to *A. lyrata*. Exons 1 and 2 showed the most diversity between *A. lyrata* and *A. thaliana* with 6 and 7 amino acid substitutions, respectively. Only one single amino acid substitution was found in exon 3. *A. thaliana* had one substitution in each of the N- and C-domains. Some of these coding polymorphisms could affect protein function and contribute to phenotypic variation.

### **Flanking and Intron Sequence Divergence:**

As expected, the flanking and intergenic regions contained many substitutions and insertions/deletions. Intron 1 was highlighted by an 18 bp insertion and an 11 bp deletion. Intron 2 was loaded with deletions ranging from 11 bps to 68 bps. The largest insertion for *A. thaliana* of 96 bps was found in the 3' flanking region. Numerous indels in *A. thaliana* ranging from 1 to 10 bps made Intron 2 and the 3' flanking region hardest to align. Intron 2 contains longest of the *A. lyrata* population inserts and the sequences of *A. thaliana* and *B. drummondii* have the most sequence loss here, ~200 bp's each. Evidence of directional selection on non-coding regions in *A. thaliana* has been shown to result in a general size loss in intergenic regions compared to *A. lyrata* (Wright et. al. 2002). Our data shows that the European samples had the greatest combined variability in introns 1 and 2 with lengths ranging from 1635 bps to 1933 bps, a difference of 298 bps. By comparison the *A. thaliana* standard had only 1376 bps, a loss of 16-29% relative to European *A. lyrata*. The data from the North American samples only showed a difference

of 1803 to 1806 bps resulting in loss in *A. thaliana* of 24%. The *A. lyrata* standard (from a North American genotype) did not possess the common insert found in our North American samples and the combined lengths of introns 1 and 2, 1645 bps, mirrored over half of the European populations resulting in a 16% loss of sequence in *A. thaliana*.

The initial sequence comparison of *A. thaliana* and *A. lyrata* found the presence of three large inserts in the intergenic regions of *A. lyrata*. These inserts formed stem loop structures typical of retrotransposons and miRNAs. Two of the inserts were only found in some of the European alleles and the other insert was only found in alleles from NA. The largest NA insert was found in all of our NA samples and was located within 65 bps of the largest European insert found only in the S5-21A population. The sequence data in the region between the NA insert and the European S5-21A insert contained numerous polymorphisms, all but one of which were also in the European populations. The sequence data also showed the European populations, except two, had an ~83 bp insert in common in the 3' flanking region relative to the *A. lyrata* standard. The *A. thaliana* sequence has a longer insert, 104 bps, in the exact same area.

### **Function of Insertions:**

I inspected the inserts for functional significance to test the hypothesis that *LFY* polymorphisms are the result of natural selection interacting with environmental conditions in different geographic locations. The 241 bp insert in the Spiterstulen allele S5-21A is part of the Little Athila Ty3/Gypsy group of LTR retrotransposons (Griffiths-Jones et al 2005/2007, Marín and Lloréns 2000). Research shows that nine of these Ty3/Gypsy lineages are present in *A. thaliana* and this one is found on all five

chromosomes of *A. thaliana* (Marín and Lloréns 2000). Even though these elements predate the monocot/eudicot split of 200 mya, it was not found in any of the NA populations (Marco and Marín 2005, 2008). It is interesting to note that a previous examination of 87 introns from *A. lyrata* did not find any evidence of such a large insert, but did find smaller ones (Wright et al 2002). The events surrounding the insertion of this retro element may have occurred in the ancestral population that migrated northward and having been largely neutral still exists in the Spiterstulen population. These elements and those like them have been hypothesized as being largely responsible for the huge increase in the *A. lyrata* genome size (~235 million base pairs) compared to *A. thaliana* (~125 million base pairs; Wright et al 2002).

The other significant European insertion consisted of ~83 bps inserted in the 3' flanking region of all the Spiterstulen samples, 2 out of 4 Plech samples, the Olsen et. al. standard sequence, and the *A. thaliana* genome sequence. This insert is located in the 3' UTR, starts upstream of a known polyadenylation site and could be co-transcribed with *LFY*. Information from miRBase shows that this insert is part of the ath-MIR 169 family that targets mRNAs coding for *CCAAT binding factor (CBF)-HAP2 like* proteins. The MIR169 family is found on all five chromosomes of *A. thaliana* and is known to interact with *CONSTANS (CO)*, a primary upstream regulator of the photoperiod pathway to floral induction (Laubinger et al 2006, Li et al 2010, Wenkel et al 2006). It may be possible that the concurrent transcription of ath-MIR169 with *LFY* works as a negative regulator of the photoperiod pathway during the rapid change in photoperiods. When *LFY* is transcribed the miRNA transcription start site may be activated. The amount of

day length that needs to be responded to between the Summer and Winter solstice varies greatly with location. The average difference between the length of sunlight exposure on the longest (June 21<sup>st</sup>) and shortest days (December 22<sup>nd</sup>) ranges from 12h: 57m: 06s in Norway to 4h: 54m: 18s in Mayodan, N.C. During the longest day, Norwegian populations will be exposed to almost 19 hours of sunlight versus 14.5 hours for Mayodan populations roughly 4.5 hours difference. By the time the winter solstice arrives the Norwegian areas receive four hours less light. Thus, the NA populations will receive longer periods of light during the beginning and end of the growing season and can thereby extend the length of time it is able to flower or grow vegetatively. The large variation in photoperiods experienced by European populations contrasts greatly with the NA populations and the European populations may require more components of the photoperiod pathway expanding its ability to respond to stimuli (Li et al 2010, Laubinger et al 2006, and Wenkel et al 2006). When the photoperiod pathway is repressed the autonomous pathway can take over. This pathway depends on other environmental cues than the photoperiod pathway and could promote *LFY* induction through one of its known promoters, AP1. This mechanism may provide a manner in which *A. lyrata* is able to fine tune its response to stimuli. This particular mechanism could be tested by performing a protein assay for the levels of CO in the shoot apex of plants with and without the insert. We would expect to find lower levels in the plants with the insert upon induction of *LFY*.

The discovery of these ath-MIR elements in *LFY* is consistent with previous findings using over 12,500 sequences that five TF-binding motifs, (LFY, TATA-box, AtMYC2-bHLH, ARF, and SORLREP3) dubbed “miRNA-preferred motifs”, are



overrepresented in miRNA promoter regions relative to the gene-coding promoter regions (Wang et al 2009). Since *LFY* has LFY binding sites these types of elements are not unexpected. The role of miRNAs in animal development has been established and is consistent with the developmental role that *LFY* plays in Arabidopsis (Papadopoulos et al 2009, <http://diana.cslab.ece.ntua.gr/tarbase/>, Megraw et al 2006). This overrepresentation of *LFY* leads to an interesting hypothesis: Induction of *LFY* transcription leads to the activation of multiple miRNAs which in turn down regulates a wide variety of genes. As the quantity of CO mRNA increases, it eventually reaches a level at which its regulatory miRNA can no longer prevent translation. Once this threshold has been passed, the available CO proteins will then induce transcription of LFY which will induce transcription of the miRNA. Since *LFY* will self promote its own transcription it will no longer need the original stimuli nor will it require the cell to expend unnecessary energy or waste resources producing that original signal. *LFY* alleles containing MIR 169 in their 3' UTRs could induce further attenuation of the photoperiod pathway.

### **Putative miRNA:**

The North American insert in intron 2, ~160 bps, was identified as ath-MIR 157 which acts like the ath-MIR 156 family and targets mRNAs for *SQUAMOSA PROMOTER BINDING PROTEIN-LIKE (SPL)* genes. The relevance here is that *SQUAMOSA* is the *API* homologue first identified in Snapdragons and these elements are responsible for up-regulating *API*. The MIR 156/157 family has been shown to target 11 out of 17 *SPL* genes in *A. thaliana*. The MIR 156 family is a known regulator of the autonomous or age-related floral induction pathway. It affects flowering time and

development through activation of *API* and *SOC1* which in turn induces transcription of *LFY* (Fornara, F., and Coupland, G. 2009, Wang et al 2009). *API* receives inductive signals from the photoperiod and age-related pathways. Down-regulation of the autonomous pathway could put more emphasis on signals received from the photoperiod pathway and allow the plant the ability to fine tune its response to stimuli. Populations that are closer to the equator will experience favorable growing conditions, besides photoperiod, for an extended period of time compared to populations farther away. Therefore, the inclusion of an autonomous element in the NA populations is logical given the longer growing season. This putative miRNA is in an intron and will be co-transcribed with the gene, and will not have to rely on an miRNA promoter in this location. This mechanism may provide a manner in which *A. lyrata* is able to fine tune its response to stimuli. This particular mechanism could be tested by performing a protein assay for the levels of the *SPL* mRNAs in the shoot apex of plants with and without the insert. We would expect to find lower levels of the *SPL* mRNAs in plants with the insert upon induction of *LFY*.

### **Potential TF Binding Sites:**

In addition to the previous insert information, evidence for other polymorphisms with possible importance in natural selection may be found by documenting the conservation or loss of putative transcription factor (TF) binding site motifs. It must be noted that these sites, unless previously proven through experimentation, have only been identified through a sequence-based search for the canonical TF motif. The binding affinity of proteins can vary with relation to sequences flanking core binding site motifs

and with variation within the core motif itself thus the need for using the canonical TF motif for our purposes. A prime example of this variation is the CArG-box motif for MADS-box proteins. The canonical site has previously only been identified as C(A/T)<sub>6</sub>G, but it now includes C(A/T)<sub>7</sub>G and C(A/T)<sub>8</sub>G variants. Additionally, the CArG motifs have reduced but not eliminated binding affinities when one of the flanking nucleotides C or G is changed (Lee et. al. 2008, Hamès et. al. 2008). Six classes of TF motifs were chosen based on their potential roles as regulators of important plant response mechanisms including hormone response and regulators of the cell cycle (ARF/Auxin, GA-Myb, and Myb), positive self-induction (LFY), pathogen infection/stress response (WRKY), and formation of the floral organs (MADS-Box) (Guo et. al. 2005, The Database of Arabidopsis Transcription Factors DATF <http://datf.cbi.pku.edu.cn/>).

Searching for ARF/Auxin TF sites, we found that *A. thaliana* had four and *A. lyrata* had two. One of the additional sites in *A. thaliana* is in an insert. The presence of these additional TF sites may allow the plant to respond in a more rapid manner to an increase in available light by binding additional ARFs. Auxin levels are up-regulated when the red to far-red ratio of the visible light spectrum shifts towards far-red and promotes cell growth and elongation. Shifts towards the red spectrum are responsible for flower initiation. Plants that are located farther away from the equator will experience much broader variation in photoperiods and at a much faster rate. These additional sites may benefit *A. thaliana* due to its semelparous life strategy. The one-time shift to flowering in *A. thaliana* may require tighter regulation of photoreceptor (phytochrome) signaling given *Arabidopsis*' preference for flowering during long daylight periods.

We chose to search for GA-Myb sites due the presence of a known binding site and their role in floral induction during short day length photoperiods. In the search for GA-Myb TF sites, all of the sequences had the known *A. thaliana* site upstream of the start codon. A search for generic MYB TF sites related to cell cycle control and regulation of the circadian clock found 4 putative sites in all of the sequences. All of the *A. lyrata* European populations except P6-12BC and P6-12AD have 1 or 2 additional sites due to inserts or polymorphisms. One would expect that having more Myb binding sites to lead to increased levels of LFY transcription under short-day conditions, and hence more flowering under those conditions. These additional sites may be related to *A. lyrata*'s iteroparous life strategy and the multiple flowering events associated with it and/or the dramatic variations in European photoperiods.

*LFY* is known to self-promote its own transcription and the search for LFY TF sites found that all of the *A. lyrata* sequences had the same six potential sites; *A. thaliana* and *B. drummondii* did not have two of these sites in intron 2 due to a base pair deletion and an insertion respectively. As mentioned before, intron 2 has been the site of considerable sequence variation. Given that LFY binds as a homodimer, the sites that are 9 bps apart are most likely true binding sites. LFY has been shown to bind to two consecutive turns of the DNA-helix as seen in Figure 20

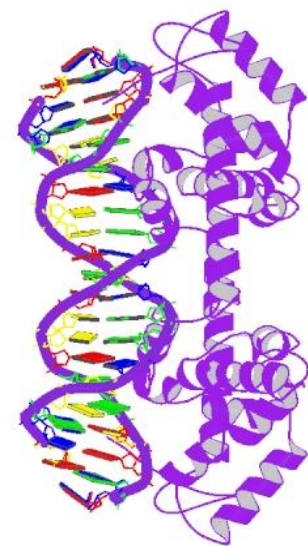


Figure 20 *LEAFY* transcription factor from *A. thaliana* in complex with DNA from *API* promoter Hamés et al 2008

(Hamés et al 2008). Since the DNA helix turns every 10 nucleotides; the 9 bases between

the end of the first binding site and the beginning of the next binding site make these very likely candidates. Conservation of these sites may be important due to the fact that this is a positive feedback mechanism and once *LFY* is being produced it will be present as long as conditions are favorable or until it is degraded. As conditions improve or reach the proper state, *LFY* will be responding to other environmental cues that will increase the levels of *LFY* and any *LFY* protein not already in a complex with another protein can continue to promote its own transcription.

All of the *Arabidopsis* sequences have 9 putative WRKY TF sites. Many WRKY proteins are considered to regulate the response to pathogen infection and other stresses as well as being involved in trichome development and the biosynthesis of secondary metabolites. *A. thaliana* has one fewer site in the 3' flanking region and gained an additional site in intron 1. The overall conservation of these putative sites would suggest that the current configuration may be required for proper pathogen/stress response and is not influenced by life-history strategy or location.

The last TF binding site motif searched for was the MADS-Box motif. The MADS-Box protein heterodimer of SOC1/AGL24 is known to be a direct upstream activator of *LFY* and SOC1/AGL24 responds to signaling from the vernalization and autonomous floral induction pathways. A known SOC1/AGL24 binding site is in the region of ~200 bps upstream of the start codon and all of the sequences retained three putative sites there (Lee et al 2008). The first 2 sites are only 8 bases apart and could be the location required; however the 2<sup>nd</sup> and 3<sup>rd</sup> sites are 108 bps apart and are mirror images of each other when compared 5'>3' and then 3'>5'. The first pair of sites may

allow a heterodimer to bind both sites at the same time on the coding strand in a linear fashion while the other set may allow binding of a heterodimer to both strands, possibly both strands at the same time.

Overall *A. thaliana* has 2 more potential binding sites than *A. lyrata*. However, but given that the MADS-Box proteins are known to function as homo- and heterodimer and forming quartets, it is difficult to allocate any significant importance to the additional sites. Functional MADS-Box dimers have been shown to only require that one dimer site be present to bind and activate a gene. Additionally, each one of the AP1/AP1, AG/AG, and AP3/PI dimers have been shown to bind three different CArG motifs with only slightly differing binding affinity (Reichmann et al 1996/1997).

#### **Population Genetics Analysis:**

The *A. lyrata* populations had polymorphisms that were shown to be geographic and population-specific. The lower levels of polymorphisms in North American populations (ours only have one codon difference total in all four samples) compared to European populations has been previously documented in other genes with populations in Plech, Germany harboring the highest levels of variation (Ross-Ibarra et. al. 2008).

This data is consistent with the Central Europe refugium scenario (Koch and Matschinger 2007, Ross-Ibarra et. al. 2008); populations radiating out from Central Europe should retain alleles found in the ancestral populations. The P3-1AC haplotype from Plech Germany had the only European amino acid sequence to exactly match the recovered sequences from the North American populations. The sequence data from my S5-21B and S9-4 alleles would be also exact matches except for a one codon deletion in

exon 3. The P3-1AC allele contained three European specific and one population specific polymorphisms; however they did not result in alterations in the amino acid sequence coded for.

It would appear that as these small *A. lyrata* populations moved across the Atlantic and south into the lower latitudes, the P3-1AC sequence was the only allelic combination not lost due to genetic drift. By contrast, as the populations moved north following the retreating ice sheets, the P3-1AC allele was one of many alleles to be retained and this may be related to the shorter distance traveled and thus less isolation of populations. These small populations will by chance only contain some of the genetic variants present in the ancestral population. The limited population numbers with reduced diversity would lead to a lower number of fertilization events and would thereby reduce gene flow. This scenario, referred to as the founder effect, is known to enhance the effects of genetic drift and favor retention of deleterious mutations within populations that would have otherwise been eliminated through natural selection. Additional evidence of the founder effect can be seen in a comparison of all of the alleles in NA; they all have exactly the same amino acid sequence except for one amino acid change in N9-1 at the end of exon 3 just beyond the end of the C-Domain. If a large number of individuals had been responsible for the colonization of NA we would expect to see similar levels of variation like that seen in the European populations. It may be possible that other alleles were not recovered when we cloned the initial populations but the evidence does not lend itself to the idea that the NA populations have a lot of allelic variation when compared to the European populations.

The population genetics analysis of the sequence fragments provides insights into evolutionary processes acting on *LFY* and this would be evident in side-by-side comparisons of the coding vs. non-coding polymorphisms. The polymorphism and divergence data comparing *A. lyrata* against *A. thaliana* and NA *A. lyrata* against European *A. lyrata* showed a higher rate of non-synonymous polymorphisms within species than non-synonymous substitution between species. The difference in rates would imply a relaxation of purifying selection within species, or possibly the presence of a class of sites under balancing selection. The Bustamante et al. (2002) study comparing *A. thaliana* against *A. lyrata* found the ratio of polymorphisms within *A. thaliana* to be 2.33 times that of the ratio of fixed polymorphisms between the two species. These figures are reflective of the ratios we found in our study where the ratio of polymorphisms within *A. lyrata* populations was 2.36 times the ratio between *A. lyrata* and *A. thaliana*. They found in a comparison of amino-acid replacements within *Arabidopsis* species (*A. thaliana* and *A. lyrata*), against those within *Drosophila* species (*D. melanogaster* and *D. simulans*), that most of the substitutions in *Drosophila* were beneficial while the ones in *Arabidopsis* were primarily detrimental. This was considered to be a result of inbreeding in *A. thaliana* and that smaller populations have difficulty getting rid of deleterious mutations due to genetic drift (Bustamante et al. 2002).

The results from the MK tests were significant for the *A. lyrata* vs. *A. thaliana* (p-value 0.004727) and the Europe vs. *A. thaliana* (p-value 0.008964). These two tests both had Neutrality Index values greater than 3.8. The tests between NA and *A. thaliana* were not considered to be significant, most likely due to the smaller sample set; NA had five



samples and Europe had 11. Our data is consistent with previous studies showing that higher rates of nonsynonymous amino acid polymorphisms compared to divergence are often found in *Arabidopsis* genes and are usually associated with a low level of selection against deleterious mutations (Flowers et. al. 2009). Our data suggest that selection is acting more on local populations and geographic regions then across the whole species. This same idea has been proposed for the phenotypic variation in *A. thaliana* and was associated with an accumulation of polymorphisms in five other genes (Flowers et. al. 2009).

Additional evidence for this interpretation was found in the result of the Tajima's D tests. The value of D was more negative for non-synonymous sites than for synonymous sites. These tests suggest that purifying selection is taking place but is not yet complete since the non-synonymous polymorphisms are skewed towards being rare and have not been eliminated yet. The NA results did not provide any significant values and may be related to the smaller sample size. The results from the Fay and Wu's H test also confirm there is no evidence for selection to favor common derived substitutions. Again this may be related to the excess of non-synonymous polymorphisms found and rare deleterious mutations not being eliminated.

The statistical analysis shows evidence for incomplete purifying selection on *LFY*. There were seven alleles represented in the European populations and only two in the North American populations. The rate of out-crossing probably limits the opportunity for homozygous deleterious recessives to occur and be selected against, thus slowing the elimination process.

The evidence collected here supports our hypotheses that phenotypic variation in *LFY* is potentially associated with polymorphisms found in the coding and non-coding regions of *LFY*. All the population genetic analyses seem to support the relaxation of purifying selection hypothesis, rather than adaptive changes. On the other hand, the potential functional importance of some indels and gain/loss of TF binding sites indicates the potential for some sequence variation to have adaptive importance. The isolation of populations during the last glacial period has been viewed as a bottleneck event and may have provided the mechanism for these deleterious mutations to be maintained at an appreciable frequency because genetic drift is amplified in such cases and can result in deleterious mutations going to fixation at a much higher rate than they would in a large population. It would appear that cost of inbreeding and the effects of genetic drift in small populations can produce similar results out of two completely different situations

Our research has added to the greater understanding of *LFY* and *Arabidopsis* in general. To enhance our understanding of the evolution of *A. lyrata*, we should follow up our studies with functional studies of correlated transcription levels of *LFY* alleles with and without the miRNA insertions with other flowering time genes. Future applications of this information could include the fine tuning of genetically modified crops given that *LEAFY* homologues can be found in all plants including mosses, ferns, gymnosperms, and angiosperms. Multiple copies of *LFY* in *A. thaliana* have been shown to increase flower/fruit production (Huala et al 1992). Inserting multiple copies of a natural *LFY* homologue *VFL* into a cash crop like *Vitis vinifera* the common wine grape or *ZFL1/2* in *Zea mays*, could be of enormous economic benefit to our state and others. Not only might

these crops provide greater yield but they would not have to carry the stigma of being altered to achieve a trait like Round-Up resistance that was previously foreign to the existing plant. The transgenic approach might be especially useful in wine grapes, since cultivars are propagated clonally, and trying to modify them by breeding would break up the favorable allelic combinations that make particular varieties useful.

## LITERATURE CITED

- Al-Shehbaz, I.A., and O'Kane, S.L. (2002). Taxonomy and Phylogeny of *Arabidopsis* (Brassicaceae). *The Arabidopsis Book*. American Society of Plant Biologists, Rockville, Md. **30**: 1-22.
- Al-Shehbaz, I.A. (2003). *Transfer of most North American species of Arabis to Boechera* (Brassicaceae). *Novon: A Journal for Botanical Nomenclature* **13**(4): 381-391.
- Arabidopsis lyrata* genome sequencing project <http://www.phytozome.net/alyrata.php> 06/16/2010
- Arabidopsis* Small RNA Project (ASRP) Oregon State University (2010) <http://asrp.cgrb.oregonstate.edu/db/microRNAfamily.html> 06/16/2010
- Battaglia, R., Brambilla, V., Colombo, L., Stuitje, A.R., and Kater, M.M. (2006). Functional analysis of MADS-box genes controlling ovule development in *Arabidopsis* using the ethanol-inducible *alc* gene-expression system. *Mechanisms of Development* **123**: 267–276.
- Beaulieu, J., Jean, M., and Belzile, F. (2007). Linkage maps for *Arabidopsis lyrata* subsp. *lyrata* and *Arabidopsis lyrata* subsp. *petraea* combining anonymous and *Arabidopsis thaliana*-derived markers. *Genome* **50**: 142-150.
- Bell, G. (1980). The Costs of Reproduction and Their Consequences. *The American Naturalist* **116**(1): 45-76.
- Benfey, P.N. and Weigel, D. (2001). Transcriptional Networks Controlling Plant Development. *Plant Physiology* **125**: 109–111.
- Blanc, G., Barakat, A., Guyot, R., Cooke, R., and Delseny, M. (2000). Extensive Duplication and Reshuffling in the *Arabidopsis* Genome. *The Plant Cell* **12**: 1093–1101.
- Blázquez, M.A., Soowall, L.N., Ilha, L. and Weigel, D. (1997). *LEAFY* expression and flower initiation in *Arabidopsis*. *Development* **124**: 3835-3844.
- Blázquez, M.A., Green, R., Nilsson, O., Sussman, M.R., and Weigel, D. (1998). Gibberellins Promote Flowering of *Arabidopsis* by Activating the *LEAFY* Promoter. *The Plant Cell* **10**: 791–800.
- Blázquez, M.A., Ferrándiz, C., Madueño, F., and Parcy, F. (2006). How floral meristems are built. *Plant Molecular Biology* **60**: 855–870.

- Bonser, S.P. and Aarssen, L.W. (2006). Meristem allocation and life-history evolution in herbaceous plants. *Canadian Journal of Botany* **84**: 143-150.
- Bowman, J.L., Smyth, D.R., and Meyerowitz, E.M. (1989). Genes Directing Flower Development in *Arabidopsis*. *The Plant Cell* **1**: 37-52.
- Busch, M.A., Bomblies, K., and Weigel, D. (1999). Activation of a Floral Homeotic Gene in *Arabidopsis*. *Science* **285**: 585-587.
- Bustamante, C.D., Nielsen, R., Sawyer, S.A., Olsen, K.M., Purugganan, M.D., and Hartl, D.L. (2002). The cost of inbreeding in *Arabidopsis*. *NATURE* **416**: 531-534
- Bürglin, T.R. (1996) Homeodomain Proteins. In Meyers, R.A. (ed.), Encyclopedia of Molecular Biology and Molecular Medicine **3**: 55-76.
- Cary, A.J., Che P., and Howell, S.H. (2002). Developmental events and shoot apical meristem gene expression patterns during shoot development in *Arabidopsis thaliana*. *The Plant Journal* **32**: 867-877.
- Chae, E., Tan, Q.K.G., Hill, T.A., and Irish, V.F. (2008). An *Arabidopsis* F-box protein acts as a transcriptional co-factor to regulate floral development. *Development* **135**: 1235-1245.
- Clauss, M.J. and Koch, M.A. (2006). Poorly known relatives of *Arabidopsis thaliana*. *Trends in Plant Science* **11**: 449-459.
- Doust, J.L. (1989). Plant resource strategies and resource allocation. *Trends in Ecology and Evolution* **4**: 230-234.
- Edwards, K.D., Anderson, P.E., Hall, A., Salathia, N.S., Locke, J.C.W., Lynn, J.R., Straume, M., Smith, J.Q., and Millar, A.J. (2006). *FLOWERING LOCUS C* mediates natural variation in the high-temperature response of the *Arabidopsis* circadian clock. *The Plant Cell* **18**: 639-650.
- Ehrenreich, I.M. and Purugganan, M.D. (2006). The molecular genetic basis of plant adaptation. *American Journal of Botany* **93**: 953-962.
- Eriksson, S., Böhlenius, H., Moritz, T., and Nilsson, O. (2006). GA4 Is the Active Gibberellin in the Regulation of *LEAFY* Transcription and *Arabidopsis* Floral Initiation. *The Plant Cell* **18**: 2172-2181.
- Fahlgrena, N., Jogdeo, S., Kasschau, K.D., Sullivan, C.M., Chapman, E.J., Laubinger, S., Smith, L.M., Dasenko, M., Givana, S.A., Weigel, D., and Carrington, J.C. (2010). MicroRNA Gene Evolution in *Arabidopsis lyrata* and *Arabidopsis thaliana*. *The Plant Cell* **22**: 1074-1089.

- Fornara, F., and Coupland, G. (2009). Plant Phase Transitions Make a SPLash. *CELL* **138**(4): 625-627.
- Gazzani, S., Gendall, A.R., Lister, C., and Dean, C. (2003). Analysis of the Molecular Basis of Flowering Time Variation in Arabidopsis Accessions. *Plant Physiology* **132**: 1107–1114.
- Greb, T., Clarenz, O., Schäfer, E., Müller, D., Herrero, R., Schmitz, G., and Theres, K. (2003). Molecular analysis of the *LATERAL SUPPRESSOR* gene in *Arabidopsis* reveals a conserved control mechanism for axillary meristem formation. *Genes & Development* **17**: 1175-1187.
- Gregis, V., Sessa, A., Colombo, L., and Kater, M.M. (2008). AGAMOUS-LIKE24 and SHORT VEGETATIVE PHASE determine floral meristem identity in Arabidopsis. *The Plant Journal* **56**: 891–902.
- Griffiths-Jones, S., Grocock, R.J., van Dongen, S., Bateman, A., and Enright, A.J. (2006). miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Research* **34**: 140-144.
- Griffiths-Jones, S., Saini, H.K., van Dongen, S., and Enright, A.J. (2007). miRBase: tools for microRNA genomics. *Nucleic Acids Research* **36**: 154-158.
- Guo, A., He, K., Liu, D., Bai, S., Gu, X., Wei, L., and Luo, J. (2005). DATF: a Database of Arabidopsis Transcription Factors. *Bioinformatics* **21**: 2568-2569.
- Hamès, C., Pchelkine, D., Grimm, C., Thevenon, E., Moyroud, E., Gérard, F., Martiel, J.L., Benlloch, R., Parcy, F., and Müller, C.W. (2008). Structural basis for *LEAFY* floral switch function and similarity with helix-turn-helix proteins. *The EMBO Journal* **27**: 2628–2637.
- Hong, R.L., Hamaguchi, L., Busch, M.A., and Weigel, D. (2003). Regulatory Elements of the Floral Homeotic Gene *AGAMOUS* Identified by Phylogenetic Footprinting and Shadowing. *The Plant Cell* **15**: 1296–1309.
- Huala, E. and Sussex, I.M. (1992). *LEAFY* interacts with floral homeotic genes to regulate *Arabidopsis* floral development. *Plant Cell* **4**: 901-913.
- Hudson, R. R., M. Kreitman, and M. Aguade. (1987). A test of neutral molecular evolution based on nucleotide data. *Genetics* **116**: 153-159.
- Jack, T. (2004). Molecular and Genetic Mechanisms of Floral Control. *The Plant Cell* **16**: 1–17.

- Jofuku, K.D., den Boer, B.G.W., Montagu, M.V., and Okamuro, J.K. (1994). Control of *Arabidopsis* Flower and Seed Development by the Homeotic Gene *APETALA2*. *The Plant Cell* **6**: 1211-1225.
- Kater, M.M., Dreni, L., and Colombo, L. (2006). Functional conservation of MADS-box factors controlling floral organ identity in rice and *Arabidopsis*. *J. of Exp. Bot.* **57**(13): 3433–3444.
- Kaufmann, K., Melzer, R., and Theigen, G. (2005). MIKC-type MADS-domain proteins: structural modularity, protein interactions and network evolution in land plants. *Gene* **347**: 183–198.
- Kipreos, E.T., and Pagano, M. (2000). The F-box protein family. *Genome Biology* **1**(5): 3002.1-3002.7.
- Koch, M.A. and Matschinger, M. (2007) Evolution and genetic differentiation among relatives of *Arabidopsis thaliana*. *PNAS* **104**(15): 6272–6277.
- Kong, H., Leebens-Mack, J., Ni, W., de Pamphilis, C.W., and Ma, H. (2004). Highly Heterogeneous Rates of Evolution in the *SKPI* Gene Family in Plants and Animals: Functional and Evolutionary Implications. *Mol. Biol. Evol.* **21**(1): 117–128.
- Koornneef, M., Alonso-Blanco, C., and Vreugdenhil, D. (2004). Naturally Occurring Genetic Variation In *Arabidopsis thaliana*. *Annu. Rev. Plant Biol.* **55**: pp. 141–172.
- Krizek, B.A. and Fletcher, J.C. (2005). Molecular Mechanisms Of Flower Development: An Armchair Guide. *Nature Reviews Genetics* **6**: 688– 698.
- Kuittinen, H., de Haan, A.A., Vogl, C., Oikarinen, S., Leppälä, J., Koch, M., Mitchell-Olds, T., Langley, C.H., and Savolainen, O. (2004). Comparing the linkage maps of the close relatives *Arabidopsis lyrata* and *A. thaliana*. *Genetics* **168**: 1575-1584.
- Laubinger, S., Marchal, V., Gentilhomme, J., Wenkel, S., Adrian, J., Jang, S., Kulajta, C., Braun, H., Coupland, G., and Hoecker, U. (2006) *Arabidopsis* SPA proteins regulate photoperiodic flowering and interact with the floral inducer CONSTANS to regulate its stability. *Development* **133**: 3213-3222.
- Laughon, A. (1991). DNA binding specificity of homeodomains. *Biochemistry* **30**(48): 11357-11367.
- Li, Y., Fu, Y., Ji, L., Wu, C., and Zheng, C. (2010) Characterization and expression analysis of the *Arabidopsis* mir169 family. *Plant Science* **178**: 271–280.
- Liang, X., Nazarenius, T.J., and Stone, J.M. (2008). Identification of a Consensus DNA-Binding Site for the *Arabidopsis thaliana* SBP Domain Transcription Factor, *AtSPL14*, and Binding Kinetics by Surface Plasmon Resonance. *Biochemistry* **47**(12): 3645-3653.

- Ma, H., Yanofsky, M.F., and Meyerowitz, E.M. (1991). *AGL1-AGL6*, an *Arabidopsis* gene family with similarity to floral homeotic and transcription factor genes. *Genes & Development* **5**: 484-495.
- Ma, H. (1994). The unfolding drama of flower development: recent results from genetic and molecular analyses. *Genes & Development* **8**: 745-756.
- Mable, B.K. and Adam, A. (2007). Patterns of genetic diversity in out-crossing and selfing populations of *Arabidopsis lyrata*. *Molecular Ecology* **16**: 3565–3580.
- Mack, G. S. (2007). MicroRNA gets down to business. *Nature Biotechnology* **25**: 631-638.
- Maizel, A., Busch, M.A., Tanahashi, T., Perkovic, J., Kato, M., Hasebe, M., and Weigel, D. (2005). The Floral Regulator *LEAFY* Evolves by Substitutions in the DNA Binding Domain. *Science* **308**: 260-263.
- Mandel, M.A., Gustafson-Brown, C., Savidge, B. and Yanofsky, M.F. (1992). Molecular characterization of the *Arabidopsis* floral homeotic gene *APETALA1*. *Nature* **360**: 273-277.
- Marín, I. and Lloréns, C. (2000). *Ty3/Gypsy* Retrotransposons: Description of New *Arabidopsis thaliana* Elements and Evolutionary Perspectives Derived from Comparative Genomic Data. *Mol. Biol. Evol.* **17**(7): 1040–1049.
- Marco, A. and Marin, I. (2005). Retrovirus-like elements in plants. *Recent Res. Devel. Plant Sci.* **3**: 1-10.
- Marco, A. and Marin, I. (2008). How Athila retrotransposons survive in the *Arabidopsis* genome. *BMC Genomics* **9**: 1-14.
- McDonald, J. H., and M. Kreitman. (1991). Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* **351**: 652-654.
- Megraw, M., Baev, V., Rusinov, V., Jensen, S.T., Kalantidis, K., and Hatzigeorgiou, A.G. (2006). MicroRNA promoter element discovery in *Arabidopsis*. *RNA* **12**: 1612-1619.
- Meiklejohn, C.D., Montooth, K.L., and Rand, D.M. (2007). Positive and negative selection on the mitochondrial genome. *TRENDS in Genetics* **23**(6): 259-263.
- Melzer, R., Verelst, W., Theißen, G. (2009). The class E floral homeotic protein SEPALLATA3 is sufficient to loop DNA in 'floral quartet'-like complexes *in vitro*. *Nucleic Acids Research* **37**(1): 144–157.



- Michaels, S.D. and Amasino, R.M. (1999). *FLOWERING LOCUS C* Encodes a Novel MADS Domain Protein That Acts as a Repressor of Flowering. *The Plant Cell* **11**: 949–956.
- Mitchell-Olds, T. (2001). *Arabidopsis thaliana* and its wild relatives: a model system for ecology and evolution. *TRENDS in Ecology & Evolution* **16**(12): 693-700.
- Møller, S.G., and Chua, N. (1999). Interactions and Intersections of Plant Signaling Pathways. *J. Mol. Biol.* **293**: 219-234.
- Nilsson, O., Lee, I., Blázquez, M.A., and Weigel, D. (1998). Flowering-Time Genes Modulate the Response to *LEAFY* Activity. *Genetics* **150**: 403–410.
- O’Kane Jr., S.L. and Al-Shehbaz, I.A. (1997). A Synopsis of *Arabidopsis* (Brassicaceae). *Novon* **7**(3): 323-327.
- Ohshima, S., Murata, M., Sakamoto, W., Ogura, Y., and Motoyoshi, F. (1997). Cloning and molecular analysis of the *Arabidopsis* gene *Terminal Flower 1*. *Molecular and General Genetics* **254**: 186-194.
- Olsen, K.M., Womack, A., Garrett, A.R., Suddith, J.I., and Purugganan, M.D. (2002). Contrasting Evolutionary Forces in the *Arabidopsis thaliana* Floral Developmental Pathway. *Genetics* **160**: 1641–1650.
- Onouchi, H., Igeño, M.I., Périlleux, C., Graves, K., and Coupland, G. (2000). Mutagenesis of Plants Overexpressing *CONSTANS* Demonstrates Novel Interactions among *Arabidopsis* Flowering-Time Genes. *The Plant Cell* **12**: 885–900.
- Österberg, M.K., Shavorskaya, O., Lascoux, M., and Lagercrantz, U. (2002). Naturally Occurring Indel Variation in the *Brassica nigra* *COL1* Gene Is Associated with Variation in Flowering Time. *Genetics* **161**: 299–306.
- Page, D.R. and Grossniklaus, U. (2002). The Art and Design of Genetic Screens: *Arabidopsis thaliana*. *Nature Reviews Genetics* **3**: 124-136.
- Papadopoulos, G.L., Reczko, M., Simossis, V.A., Sethupathy, P., and Hatzigeorgiou, A.G. (2009) The database of experimentally supported targets: a functional update of TarBase. *Nucleic Acids Res.* **37**: 155-158.
- Parcy, F., Bomblies, K., and Weigel, D. (2002). Interaction of *LEAFY*, *AGAMOUS* and *TERMINAL FLOWER1* in maintaining floral meristem identity in *Arabidopsis*. *Development* **129**: 2519-2527.
- Reichmann, J., Krizek, B., and Meyerowitz, E. (1996). Dimerization specificity of *Arabidopsis* MADS domain homeotic proteins APETALA1, APETALA3, PISTILLATA, and AGAMOUS. *Proc. Natl. Acad. Sci. USA* **93**: 4793-4798.

- Remington, D.L. and Purugganan, M.D. (2003). Candidate Genes, Quantitative Trait Loci, and Functional Trait Evolution in Plants. *Int. J. Plant Sci.* **164**(3): 7–20.
- Riihimäki, M. and Savolainen, O. (2004). Environmental and Genetic Effects on Flowering Differences between Northern and Southern Populations of *Arabidopsis lyrata* (Brassicaceae). *American Journal of Botany* **91**(7): 1036–1045.
- Riihimäki, M., Podolsky, R., Kuittinen, H., Koelewijn, H., and Savolainen, O. (2005). Studying genetics of adaptive variation in model organisms: flowering time variation in *Arabidopsis lyrata*. *Genetica* **123**: 63–74.
- Ross-Ibarra, J., Wright, S.I., Foxe, J.P., Kawabe, A., DeRose-Wilson, L., Gos, G., Charlesworth, D., and Gaut, B.S. (2008). Patterns of Polymorphism and Demographic History in Natural Populations of *Arabidopsis lyrata*. *PLoS ONE*. **3**(6): e2411.
- Saddic, L.A., Huvermann, B., Bezhani, S., Su, Y., Winter, C.M., Kwon, C.S., Collum, R.P., and Wagner, D. (2006). The LEAFY target *LMII* is a meristem identity regulator and acts together with LEAFY to regulate expression of *CAULIFLOWER*. *Development* **133**: 1673-1682.
- Samach, A. (1999). The *UNUSUAL FLORAL ORGANS* gene of *Arabidopsis thaliana* is an F-box protein required for normal patterning and growth in the floral meristem. *The Plant Journal* **20**(4): 433-445.
- Samach, A., Onouchi, H., Gold, S.E., Ditta, G.S., Schwarz-Sommer, Z., Yanofsky, M.F., and Coupland, G. (2000). Distinct Roles of CONSTANS Target Genes in Reproductive Development of *Arabidopsis*. *Science* **288**: 1613-1616.
- Schmid, M., Uhlenhaut, N.H., Godard, F., Demar, M., Bressan, R., Weigel, D., and Lohmann, J.U. (2003). Dissection of floral induction pathways using global expression analysis. *Development* **130**: 6001-6012.
- Schranz, M.E., Lysak, M.A., and Mitchell-Olds, T. (2006). The ABC's of comparative genomics in the Brassicaceae: building blocks of crucifer genomes. *TRENDS in Plant Science* **11**(11): 535-542.
- Schultz, E.A. and Haughn, G.W. (1991). *LEAFY*, a Homeotic Gene That Regulates Inflorescence Development in *Arabidopsis*. *The Plant Cell* **3**: 771-781.
- Sharma, V.K. and Fletcher, J.C. (2002). Maintenance of Shoot and Floral Meristem Cell Proliferation and Fate. *Plant Physiology* **129**: 31–39.
- Shepard, K.A. (2007). The molecular population genetics of shoot development in *Arabidopsis thaliana*. *Genetica* **129**: 19–36.

Simpson, G.G. and Dean, C. (2002). *Arabidopsis*, the Rosetta Stone of Flowering Time? *Science* **296**: 285-289.

Sung, S. and Amasino, R.M. (2005). Remembering Winter: Toward a Molecular Understanding of Vernalization. *Annual Review of Plant Biology* **56**: 491–508.

TarBase v.5c DIANA – DNA Intelligent Analysis Tarbase Web Server (2010)  
<http://diana.cslab.ece.ntua.gr/tarbase/> 06-16-2010

Thornsberry, J.M., Goodman, M.M., Doebley, J., Kresovich, S., Nielsen, D., and Buckler, E.S. (2001). *Dwarf8* polymorphisms associate with variation in flowering time. *Nature Genetics* **28**: 286-289.

Vorobiev, V.A., Martynov, V.V., Pankin, A.A. and Khavkin, E.E. (2005). Polymorphism of the *LEAFY* Gene in Brassica Plants. *Russian Journal of Plant Physiology* **52**(6): 814–820.

Wang, J.W., Czech, B., and Weigel, D. (2009). miR156-Regulated SPL Transcription Factors Define an Endogenous Flowering Pathway in *Arabidopsis thaliana*. *Cell* **138**(4): 738-749.

Wang, X., Zhang, Y., Ma, Q., Zhang, Z., Xue, Y., Bao, S., and Chong, K. (2007). SKB1-mediated symmetric dimethylation of histone H4R3 controls flowering time in *Arabidopsis*. *The EMBO Journal* **26**: 1934–1941.

Weigel, D. (1995). The *APETALA2* Domain Is Related to a Novel Type of DNA Binding Domain. *Plant Cell* **7**: 388-389.

Wenkel, S., Turck, F., Singer, K., Gissot, L., Le Gourrierc, J., Samach, A., and Coupland, G. (2006) CONSTANS and the CCAAT Box Binding Complex Share a Functionally Important Domain and Interact to Regulate Flowering of *Arabidopsis*. *The Plant Cell* **18**: 2971–2984.

William, D.A., Su, Y., Smith, M.R., Lu, M., Baldwin, D.A., and Wagner, D. (2004). Genomic identification of direct target genes of *LEAFY*. *PNAS* **1**(6): 1775–1780.

Wright, S. I., Lauga, B., and Charlesworth, D. (2002). Rates and Patterns of Molecular Evolution in Inbred and Outbred *Arabidopsis*. *Molecular Biology and Evolution* **19**: 1407-1420.

## APPENDIX

### Appendix A.1 Region lengths in Sequence Standards and Alleles Amplified

		Region covered and Length in bp		InDels < 10 bp are not referenced here (too many).					
Sample Group	Sample ID	5' flanking	InDels > 10 bp	EXON	InDels + - / (Subst)	INTRON	InDels > 10 bp	EXON 2	InDels + - / (Subst)
Alyr/NA/Standard	Alyr Trace	514 bp	standard	457 bp	standard	462 bp	standard	401 bp	standard
Alyr/North America	N1-3	514	no change	457	(1)	462		401	
Alyr/North America	N9-1	501	-12	457	(1)	462		401	
Alyr/North America	N10-14	509		457	(1)	462		401	
Alyr/North America	I1-1	503	-10	457	(1)	472	+10	401	
Alyr/Origin Unknown	Olsen et al	496	-18,	457	(6)	541	+88	401	(2)
Alyr/Europe	P6-12BC	493	-21	457	(2)	535	+56	401	(2)
Alyr/Europe	P6-12AD	493	-21	457	(2)	535	+56	401	
Alyr/Europe	P3-1AC	491	-21	457	(2)	462		401	(1)
Alyr/Europe	P3-1BD	495	-21	457	(2)	464		401	(4)
Alyr/Europe	S1-1AC	494	-21	457	(3)	464		401	(6)
Alyr/Europe	S1-1BD	495	-21	457	(2)	464		401	(6)
Alyr/Europe	S5-21A	493	-21	457	(2)	481		401	
Alyr/Europe	S5-21B	493	-21	457	(2)	481		401	
Alyr/Europe	S9-4	493	-21	457	(3)	479		401	
Alyr/Europe	S10-4ABD	498	-16	457	(2)	462		401	(2)
Alyr/Europe	S10-4C	498	-16	457	(2)	462		401	(2)
Athaliana/Outgroup	Athaliana	555	+36, -19	457	(18)	470	-11, +18	395	-6, (16)
Boechera/Outgroup	B. drummondii			430	(20)	465	+17, +13, -11	392	-3, (22)
							-20		

## Appendix A.2 Region lengths in Sequence Standards and Alleles Amplified-cont'd

		InDels < 10 bp are not referenced here (too many).						
Sample Group	Sample ID	INTRON 2	InDels > 10 bp	EXON 3	InDels + - / (Subst)	3' flanking	InDels > 10 bp	Total length
Alyr/NA/Standard	Alyr Trace	1183 bp	standard	411 bp	standard	320 bp	standard	3748 bp
Alyr/North America	N1-3	1344	+161	411		320		3909
Alyr/North America	N9-1	1341	+158	411	(1)	320		3893
Alyr/North America	N10-14	1344	+161	411		320		3904
Alyr/North America	I1-1	1344	+161	411	(1)	320		3908
Alyr/Origin Unknown	Olsen et al	1176		411	(1)	202	+67	3684
Alyr/Europe	P6-12B C	1166		423	+12, (3)	320		3795
Alyr/Europe	P6-12AD	1167		423	+12, (5)	320		3796
Alyr/Europe	P3-1AC	1187		411	(2)	400	+83	3809
Alyr/Europe	P3-1BD	1183		411	(1)	400	+83	3811
Alyr/Europe	S1-1AC	1184		411	(1)	400	+83	3811
Alyr/Europe	S1-1BD	1184		411	(1)	400	+83	3812
Alyr/Europe	S5-21A	1452	+39, +242	411	(2)	400	+83	4095
Alyr/Europe	S5-21B	1178		408	-3	438	+38, +83	3856
Alyr/Europe	S9-4	1188		408	-3	439	+39, +83	3865
Alyr/Europe	S10-4ABD	1173		411	(2)	384	-12, +80	3786
Alyr/Europe	S10-4C	1178		408	-3, (1)	441	+41, +83	3845
Athaliana/Outgroup	Athaliana	906	+16, -32, -14, -57, -11, -68, -12, -29, -20	411	(18)	393	+96	3587
							+11	
Boechera/Outgroup	B. drummondii	1175	+15, -29, -12, +124, -99, -32, -11, +12, +14, +32	355	(15)			2817

### Appendix A.3 Amino acids 1-140 for all sequences

Results		10	20	30	40	50	60	70
	Alyr Trace	MDPEGFTSGL	FRWNPTRAMV	AAPPPVPPPP	QQQPATPQMA	AFGMRLGGLE	GLFGAYGIRF	YTAAKIAELG
exon 1	N1-3	.....	.....	.....	.....T.	.....	.....	.....
aa 1-152	N9-1	.....	.....	.....	.....T.	.....	.....	.....
	N10-14	.....	.....	.....	.....T.	.....	.....	.....
N-Domain	I1-1	.....	.....	.....	.....T.	.....	.....	.....
Aa 46-121	P6-12BC	.....	.....	.....	.....T.	.....	.....	.....
	P6-12AD	.....	.....	.....	.....T.	.....	.....	.....
	P3-1AC	.....	.....	.....	.....T.	.....	.....	.....
exon 2	P3-1BD	.....	.....	.....	.....T.	.....	.....	.....
Aa 153-286	S1-1AC	.....	.....	.....	.....T.	.....	.....	.....
	S1-1BD	.....	.....	.....	.....T.	.....	.....	.....
	S5-21A	.....	.....	.....	.....T.	.....	.....	.....
C-Domain	S5-21B	.....	.....	.....	.....T.	.....	.....	.....
Aa 231-390	S9-4	.....	.....	.....	.....T.	.....	.....	.....
	S10-4ABD	.....	.....	.....	.....T.	.....	.....	.....
exon 3	S10-4C	.....	.....	.....	.....A.T.	.....	.....	.....
287-422	Athaliana	.....	.....L.	.....Q.	.....L	.....V.	.....T.	.....P.
	Bdrummondii	-----	.....A.	.....Q.	.....	.....	.....T.	.....
	Olsen et al	.....	.....	.....Q.	.....	.....TR	.....	.....
Binding								
Mutation	Alyr Trace	FTASTLVGMK	DEELEEMMNS	LSHIFRWELL	VGERYGIKAA	VRAERRRLQE	EEEEESSRRR	HLLLSAAGDS
dimer ↓	N1-3	.....	.....	.....	.....	.....	.....	.....
	N9-1	.....	.....	.....	.....	.....	.....	.....
	N10-14	.....	.....	.....	.....	.....	.....	.....
	I1-1	.....	.....	.....	.....	.....	.....	.....
	P6-12BC	.....	.....	.....	.....	.....	.....	.....
	P6-12AD	.....	.....	.....	.....	.....	.....	.....
	P3-1AC	.....	.....	.....	.....	.....	.....	.....
	P3-1BD	.....	.....	.....	.....	.....	.....	.....
	S1-1AC	.....	.....	.....	.....	.....	.....	.....
	S1-1BD	.....	.....	.....	.....	.....	.....	.....
	S5-21A	.....	.....	.....	.....	.....	.....	.....
	S5-21B	.....	.....	.....	.....	.....	.....	.....
	S9-4	.....	.....	.....	.....	.....	.....	.....
	S10-4ABD	.....	.....	.....	.....	.....	.....	.....
	S10-4C	.....	.....	.....	.....	.....	.....	.....
	Athaliana	.....	.....	.....	.....	.....	.....	.....
	Bdrummondii	.....	.....	.....	.....	.....D.	.....	.....
	Olsen et al	.....	.....	.....	.....	.....T.	.....	.....

## Appendix A.4 Amino acids 141-280 for all sequences

Results		150	160	170	180	190	200	210
exon 1 aa 1-152	Alyr Trace	GTHHALDALS	QEGLSSEPVQ	QQNQTDAAAGN	NGGGGSGYWE	AGQAKMKKQQ	QQRRRKKPMV	TSVETDDVDN
	N1-3	.....	.....	.....	.....	.....	.....	.....
	N9-1	.....	.....	.....	.....	.....	.....	.....
	N10-14	.....	.....	.....	.....	.....	.....	.....
	I1-1	.....	.....	.....	.....	.....	.....	.....
	P6-12BC	.....	.....	.....	.....	.....	.....	.....
	P6-12AD	.....	.....	.....	.....	.....	.....	.....
	P3-1AC	.....	.....	.....	.....	.....	.....	.....
	P3-1BD	.....	.....	R.....	.....	.....	.....	.....
	S1-1AC	.....	.....	R.....	.....	.....	I.....	.....
exon 2 Aa 153-286	S1-1BD	.....	.....	R.....	.....	.....	.....	.....
	S5-21A	.....	.....	.....	.....	.....	.....	.....
	S5-21B	.....	.....	.....	.....	.....	.....	.....
	S9-4	.....	.....	.....	.....	.....	.....	.....
	S10-4ABD	.....	.....	.....	.....	R.....	.....	.....
exon 3 287-422	S10-4C	.....	.....	.....	.....	.....	G.....	.....
	Athaliana	.....	.....	D.....	.....	D.....	G.....	L.....E.....
	Bdrummondii	.....	.....	D.-N.....	.....	.....	.....	A.....
	Olsen et al	.....	.....	.....	.....	.....	.....	.....
		.....	.....	.....	.....	.....	.....	.....
Binding		220	230	240	250	260	270	280
Mutation dimer ↓	Alyr Trace	EGDDDDGMDN	GNGGGGGGLG	TERQREHPFI	VTEPGGEVARG	KKNGLDYLFH	LYEQCREFL	QVQTIKDRG
	N1-3	.....	.....	.....	.....	.....	.....	.....
	N9-1	.....	.....	.....	.....	.....	.....	.....
	N10-14	.....	.....	.....	.....	.....	.....	.....
	I1-1	.....	.....	.....	.....	.....	.....	.....
	P6-12BC	.....	.....S.....	.....	.....	.....	.....	.....
	P6-12AD	.....	.....	.....	.....	.....	.....	.....
	P3-1AC	.....	.....	.....	.....	.....	.....	.....
	P3-1BD	.....	.....	.....	.....	.....	.....	.....
	S1-1AC	.....	.....	.....	.....	.....	.....R.....	.....
	S1-1BD	.....	.....	.....	.....	.....	.....H.....	.....
	S5-21A	.....	.....	.....	.....	.....	.....	.....
	S5-21B	.....	.....	.....	.....	.....	.....	.....
	S9-4	.....	.....	.....	.....	.....	.....	.....
	S10-4ABD	.....	.....	.....	.....	.....	.....	.....
	S10-4C	.....	.....	.....	.....	.....	.....	.....
	Athaliana	..E.....	..--S..	.....	.....	.....	.....	.....
	Bdrummondii	..E.....	..--V..	.....	.....	.....	.....	.....
	Olsen et al	.....	.....	.....	.....	.....	.....	.....
		.....	.....	.....	.....	.....	.....	.....

# Appendix A.5 Amino acids 281-422 for all sequences

Results		290	300	310	320	330	340	350	
		..... ..... ..... ..... ..... ..... ..... ..... .....							
Alyr Trace		EKCP	TKVTNQ	VFRYAKKSGA	SYINKPKMRH	YVHCYALHCL	DEDASNALRR	AFKERGENVG	SWRQACYKPL
exon 1	N1-3	..... ..... ..... ..... ..... ..... ..... ..... .....							
aa 1-152	N9-1	..... ..... ..... ..... ..... ..... ..... ..... .....							
	N10-14	..... ..... ..... ..... ..... ..... ..... ..... .....							
	I1-1	..... ..... ..... ..... ..... ..... ..... ..... .....							
N-Domain	P6-12BC	..... ..... ..... ..... ..... ..... ..... ..... .....					S.		
Aa 46-121	P6-12AD	..... ..... ..... ..... ..... ..... ..... ..... .....			C.		S.		
	P3-1AC	..... ..... ..... ..... ..... ..... ..... ..... .....							
	P3-1BD	..... ..... ..... ..... ..... ..... ..... ..... .....							
exon 2	S1-1AC	..... ..... ..... ..... ..... ..... ..... ..... .....							
Aa 153-286	S1-1BD	..... ..... ..... ..... ..... ..... ..... ..... .....							
	S5-21A	..... ..... ..... ..... ..... ..... ..... ..... .....			T.				
C-Domain	S5-21B	..... ..... ..... ..... ..... ..... ..... ..... .....							
Aa 231-390	S9-4	..... ..... ..... ..... ..... ..... ..... ..... .....							
	S10-4ABD	..... ..... ..... ..... ..... ..... ..... ..... .....					R.		
exon 3	S10-4C	..... ..... ..... ..... ..... ..... ..... ..... .....							
287-422	Athaliana	..... ..... ..... ..... ..... ..... ..... ..... .....				E.			
	Bdrummondii	..... ..... ..... ..... ..... ..... ..... ..... .....							
	Olsen et al	..... ..... ..... ..... ..... ..... ..... ..... .....							
Binding									
		..... ..... ..... ..... ..... ..... ..... ..... .....							
Mutation		..... ..... ..... ..... ..... ..... ..... ..... .....							
		..... ..... ..... ..... ..... ..... ..... ..... .....							
dimer		..... ..... ..... ..... ..... ..... ..... ..... .....							
Alyr Trace		VNIACRHGWD	IDAVFNAHPR	LSIIVVPTKL	RQLCHLERNN	AVAAAAA---	LVGGISCTG	SSTSGRGGCG	GDDLRF*
	N1-3	..... ..... ..... ..... ..... ..... ..... ..... .....				---	---	---	*
	N9-1	..... ..... ..... ..... ..... ..... ..... ..... .....				T.	---	---	*
	N10-14	..... ..... ..... ..... ..... ..... ..... ..... .....				---	---	---	*
	I1-1	..... ..... ..... ..... ..... ..... ..... ..... .....				---	---	---	*
	P6-12BC	..... ..... ..... ..... ..... ..... ..... ..... .....				SAA	A.	---	*
	P6-12AD	..... ..... ..... ..... ..... ..... ..... ..... .....				V.	SAA	A.	*
	P3-1AC	..... ..... ..... ..... ..... ..... ..... ..... .....				---	---	---	*
	P3-1BD	..... ..... ..... ..... ..... ..... ..... ..... .....				---	---	---	*
	S1-1AC	..... ..... ..... ..... ..... ..... ..... ..... .....				---	---	---	*
	S1-1BD	..... ..... ..... ..... ..... ..... ..... ..... .....				---	---	---	*
	S5-21A	..... ..... ..... ..... ..... ..... ..... ..... .....				W.	---	---	*
	S5-21B	..... ..... ..... ..... ..... ..... ..... ..... .....				---	---	---	*
	S9-4	..... ..... ..... ..... ..... ..... ..... ..... .....				---	---	---	*
	S10-4ABD	..... ..... ..... ..... ..... ..... ..... ..... .....				---	---	---	*
	S10-4C	..... ..... ..... ..... ..... ..... ..... ..... .....				---	---	---	*
	Athaliana	..... ..... ..... ..... ..... ..... ..... ..... .....				---	---	---	*
	Bdrummondii	..... ..... ..... ..... ..... ..... ..... ..... .....				---	X-	-----	*
	Olsen et al	..... ..... ..... ..... ..... ..... ..... ..... .....				---	---	-----	*